Linear Fuzzy Clustering With Selection of Variables Using Graded Possibilistic Approach

Katsuhiro Honda, *Member, IEEE,* Hidetomo Ichihashi, *Member, IEEE,* Francesco Masulli, *Senior Member, IEEE,* and Stefano Rovetta

***Abstract*— Linear fuzzy clustering is a useful tool for knowledge discovery in databases (KDD), and several modifications have been proposed in order to analyze real world data. This paper proposes a new approach for estimating local linear models, in which linear fuzzy clustering is performed by selecting variables that are useful for extracting correlation structure in each cluster. The new clustering model uses two types of memberships. One is the conventional membership that represents the degree of membership of each sample in each cluster. The other is the additional parameter that represents the relative responsibility of each variable for estimation of local linear models. The additional membership takes large values when the variable has close relationship with local principal components, and is calculated by using the graded possibilistic approach. Numerical experiments demonstrate that the proposed method is useful for identifying local linear model taking typicality of each variable into account.**

***Index Terms*— Fuzzy clustering, principal component analysis, data mining, possibilistic clustering, variable selection.**

I. Introduction

ONE of the useful approaches to knowledge discovery in databases (KDD) is to capture the data structure of high dimensional data sets by constructing low dimensional feature space. Local principal component analysis (Local PCA) [1], [2] is an extension of linear PCA for estimating local linear sub-models by partitioning a data set into several groups. The task can be regarded as the simultaneous application of linear PCA and clustering. Linear fuzzy clustering [3], [4], [5] is a technique for partitioning samples into several linear clusters in order to capture local linear structure. In the Fuzzy c-Varieties (FCV) clustering [3], [4], each cluster is represented by its prototypical linear variety and the clustering criterion is the sum of distances between data samples and prototypes. Although the goal of linear fuzzy clustering is to capture the cluster structures of data sets, the algorithm is often identified with a technique for Local PCA because the vectors spanning the prototypical linear varieties are derived by solving the eigenvalue problems of fuzzy scatter matrices [6].

In the analysis of large scale data sets, it is often the case that they include unnecessary variables which is not informative for modeling. There are two approaches for estimating the responsibility of variables in PCA. In the dimension reduction tasks, the goal is to select items or variables so as to keep the original information as well as possible. When several variables are mutually dependent, we can estimate the data substructure even if we eliminate some of the redundant variables. In order to select the redundant variables, several criteria for variable selection in PCA have been proposed [7], [8], [9]. When we want to obtain the best subset of variables, we should search for the subset which has the largest (or smallest) criterion value among all possible subsets. On the other hand, in the data mining tasks, the goal is to extract association rules and we wish to select the variables that are mutually dependent, i.e., we should eliminate variables that have no responsibility for the estimation of principal subspace.

This paper proposes a new approach to variable selection in linear fuzzy clustering, which selects the variables that have close relationship with local principal components, by introducing the mechanism of variable selection into the iterative algorithm of linear fuzzy clustering. In order to evaluate the responsibility of variables, the proposed method uses the least squares criterion for linear fuzzy clustering [10], [11], [12], which is used for partitioning a data set based on the lower rank approximation of data matrix in each cluster. While the minimization of the least squares criterion is regarded as the component-wise approximation of a data matrix, the derived cluster structure is the same as that of the conventional FCV clustering whose clustering criteria are the squared distances between data points and prototypical linear subspaces. Several extended algorithms for linear fuzzy clustering, which can handle component-wise problems such as missing values or intra-sample outliers, have been proposed by using the least squares criterion. This paper proposes a new linear fuzzy clustering algorithm that uses two types of memberships for partitioning both samples and variables. One is the conventional membership that represents the degree of membership of each sample to each cluster. The other is the additional parameter that represents the responsibility of each variable for local model estimation in each cluster. Local subspaces are estimated emphasizing meaningful variables that have large memberships.

By the way, for variable selection in the k-Means clustering, Huang *et al.* [13] introduced the memberships of variables that play a role for fuzzy partitioning of variables. While they considered the weights for evaluating the relative importance of variables in data partitioning process, the proposed method considers the weights for evaluating the absolute importance of variables in local linear model estimation in each cluster. The new memberships measure the degree of importance of each variable using the reconstruction errors in the lower rank

Manuscript received July 21, 2005.

K. Honda and H. Ichihashi are with the Department of Computer Sciences and Intelligent Systems, Osaka Prefecture University, 1-1 Gakuen-cho, Nakaku, Sakai, Osaka 599-8531 Japan (e-mail: {honda, ichi}@cs.osakafu-u.ac.jp)

F. Masulli and S. Rovetta are with the Department of Computer and Information Sciences, University of Genova, Via Dodecaneso 35, I-16146 Genova, Italy (e-mail: {masulli, ste}@disi.unige.it)

approximation, i.e., the new measure is responsible for estimating the prediction ability in reduced-rank prediction. So, large memberships mean that the variables are well reconstructed by the local linear models, and the linear models are estimated so that they keep the information of the meaningful variables as well as possible.

When we optimize the objective function including two types of memberships, we should impose a different constraint on the additional one. As the sum of the conventional memberships over all clusters is constrained to be 1, the sum of the responsibility over all variables is to be 1. However, when the number of variables is large, the responsibility values become very small and make it difficult to interpret the absolute responsibility of the variables. To cope with this problem in the mixed c-means clustering [14], Pal *et al.* [15] proposed to relax the constraint (row sum = 1) on the typicality values but retain the column constraint on the membership values so that the additional values represent the absolute typicalities. In this paper, the graded possibilistic approach [16], [17] is applied to the estimation of memberships of variables. Because the membership of each variable represents the possibility of responsibility, they can be meaningful criteria for the variable selection in PCA.

Section II presents a brief review of the formulation of linear fuzzy clustering that uses the least squares criterion for component-wise approximation. In Section III, a technique for variable selection in linear fuzzy clustering is proposed by introducing the memberships of variables into the least squares criterion. Several numerical examples are presented in Section IV, which reveals the characteristic features of the proposed method. The final section summarizes the results of this paper.

II. Local Principal Component Analysis and Linear Fuzzy Clustering

A. Principal Component Analysis and Variable Selection

In the analysis of large scale data sets, it is often the case that they include unnecessary variables which have no useful information for modeling, i.e., we can extract data substructures, in which several variables have mutual dependencies, by eliminating unnecessary variables. In order to delete the redundant variables, several criteria for variable selection in PCA have been proposed [7], [8], [9], in which the goal is to select items or variables so as to keep the original information as well as possible. Assume that we have a 3-D data set shown in Fig. 1, in which x_1 and x_2 are mutually dependent while x_3 is almost random. If the goal is the data compression, we should calculate principal components by estimating the principal 2-D subspace, and x_1 and x_2 are redundant in estimating the substructure. So the conventional variable selection criteria tries to eliminate x_1 or x_2. When we want to obtain the best subset of variables, we should search for the subset which has the largest (or smallest) criterion value among all possible subsets. From the aspect of practical application, the one-variable stepwise procedures such as Backward elimination, Forward selection, Backward-forward stepwise selection and Forward-backward stepwise selection are used [9].

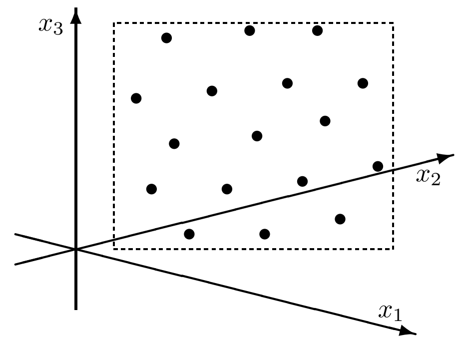


Fig. 1. Example of variable selection in PCA

On the other hand, in the data mining tasks, the goal is to extract association rules and we wish to select the variables that are mutually dependent, i.e., we should eliminate variables that can be regarded as random. In Fig. 1, x_3 seems not to have information for constructing any association rule, and we should emphasize the mutual dependency between x_1 and x_2 by estimating the linear structure (1-D principal line). In this sense, it is useful for data mining to evaluate the relative responsibilities of variables for estimation of principal subspace. The remaining part of this paper proposes a technique for estimating local linear models considering the responsibility weights of variables after a brief review of the related clustering methods.

B. Linear Fuzzy Clustering with Least Squares Criterion

Assume that we have an $(n \times m)$ data matrix $X = (x_{ij})$ consisting of m-dimensional observation of n samples. In the following, the data matrix is also represented as $X = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m)$ using n-dimensional column vectors $\boldsymbol{x}_j$'s composed of the elements of the j-th columns of X, or $X = (\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_n)^\top$ using m-dimensional column vectors $\tilde{\boldsymbol{x}}_i$'s composed of the i-th row elements of X, respectively. (All bold symbols represent column vectors and the vectors composed of the row elements of a matrix are superscripted by "$\sim$".)

Linear fuzzy clustering is a technique for partitioning samples into several linear clusters in order to capture local linear structure. Fuzzy c-Varieties (FCV) [3], [4] partitions a data set into C linear fuzzy clusters using linear varieties as the prototypes of clusters. The clustering criterion for the FCV clustering is the within-group-sum-of-errors from p-dimensional prototypical linear varieties spanned by linearly independent vectors $\boldsymbol{a}_{ck}$'s. The objective function is the weighted sum of the distances as follows:

$$L_{fcv} = \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci}^{\theta} \Big\{ (\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c)^\top (\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c) - \sum_{k=1}^{p} \boldsymbol{a}_{ck}^\top R_{ci} \boldsymbol{a}_{ck} \Big\}, \quad (1)$$

$$R_{ci} = (\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c)(\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c)^\top, \quad (2)$$

where u_{ci} denotes the membership degree of data point $\tilde{\boldsymbol{x}}_i$ to the c-th cluster and $\top$ represents the transpose of the vector (or matrix). $\boldsymbol{b}_c$ is the center of the c-th cluster. The weighting exponent θ is called "fuzzifier" and plays a role for fuzzification of memberships. The larger the θ, the fuzzier the membership assignments.

The clustering algorithm is based on the alternate optimization technique. The optimal $\boldsymbol{a}_{ck}$'s are derived from necessary condition for the optimality $\partial L_{fcv}/\partial \boldsymbol{a}_{ck} = \mathbf{0}$, and is the solution of the eigenvalue problem of fuzzy scatter matrix. Because the optimal $\boldsymbol{a}_{ck}$'s are the eigenvectors corresponding to the largest eigenvalues, the vectors can be regarded as the fuzzy principal component vectors extracted in each cluster considering memberships [6]. Consequently, cluster centers and memberships are updated from conditions $\partial L_{fcv}/\partial \boldsymbol{b}_c = \mathbf{0}$ and $\partial L_{fcv}/\partial u_{ci} = 0$, respectively.

Although the FCV clustering is useful for partitioning samples, is not suited to evaluation of the responsibilities of variables because the clustering criterion can be used only for calculating the membership degree of each sample. Another formulation of the FCV clustering have been proposed by modifying the least squares criterion for principal component analysis [18]. Introducing memberships u_{ci}'s, the least squares criterion for local PCA is defined as follows [11]:

$$\begin{aligned} L_{lsc} &= \sum_{c=1}^{C} \mathrm{tr}\Big\{(X - Y_c)^\top U_c^\theta (X - Y_c)\Big\} \\ &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}^\theta \sum_{j=1}^{m}\Big(x_{ij} - \sum_{k=1}^{p} f_{cik} a_{cjk} - b_{cj}\Big)^2, \end{aligned} \tag{3}$$

where U_c is the $(n \times n)$ diagonal matrix $U_c = \mathrm{diag}(u_{c1}, \cdots, u_{cn})$. $Y_c = (y_{cij})$ is the lower rank approximation of data matrix X, which is estimated in the c-th cluster as follows:

$$Y_c = F_c A_c^\top + \mathbf{1}_n \boldsymbol{b}_c^\top, \tag{4}$$

where $F_c = (\tilde{\boldsymbol{f}}_{c1}, \cdots, \tilde{\boldsymbol{f}}_{cn})^\top$ is the $(n \times p)$ score matrix and $A_c = (\boldsymbol{a}_{c1}, \cdots, \boldsymbol{a}_{cp})$ is the $(m \times p)$ principal component matrix. $\mathbf{1}_n$ is n-dimensional vector whose all elements are 1. Under the constraints that $F_c^\top U_c^\theta \mathbf{1}_n = \mathbf{0}$ and $A_c^\top A_c = I$, Eq.(3) is equivalent to the objective function of FCV and the minimization problem is solved by computing the p largest eigenvalues of the fuzzy scatter matrix and their associated vectors.

C. Robust Linear Fuzzy Clustering

Using the component-wise formulation, we can handle missing values and intra-sample noise. Honda and Ichihashi [10], [11], [12] introduced additional weight parameters w_{cij}'s into the least squares criterion and proposed the following objective function.

$$\begin{aligned} L_{rfcv} = &\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci} \sum_{j=1}^{m} w_{cij}\Big(x_{ij} - \sum_{k=1}^{p} f_{cik} a_{cjk} - b_{cj}\Big)^2 \\ &+\lambda \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci}, \end{aligned} \tag{5}$$

where w_{cij} is the weight corresponding to element of data matrix x_{ij}. In Eq.(5), entropy regularization approach [19] was applied for fuzzification of memberships and the entropy term was added instead of the weighting exponent in the standard FCV algorithm. The larger the λ, the fuzzier the membership assignments. (The fuzzification technique derives a similar algorithm to that of entropy-constrained fuzzy clustering by deterministic annealing (DA) [20]. As is mentioned in [21], the entropy regularized linear fuzzy clustering algorithm can also be identified with the soft version local PCA [2].) Additional weight w_{cij} represents the responsibility of element x_{ij} for local model estimation. When x_{ij} is missing, w_{cij} is set to be zero and the corresponding error is ignored [11]. When x_{ij} is regarded as noise observation in the c-th cluster, a small value is assigned to w_{cij} and the corresponding error is ignored [12]. Usually, the weights for the robust approach are calculated based on the iteratively reweighted least squares (IRLS) technique [22].

D. Hybrid Approaches to FCM-type Fuzzy Clustering

While the IRLS technique is a useful approach to robust M-estimation, the algorithm has close relationship with the robust fuzzy clustering [23], and it can be said that weight w_{cij} plays a role of memberships in robust clustering with possibilistic constraint [24]. In this sense, the robust FCV algorithm is a hybrid model of FCM-type fuzzy clustering.

Several hybrid approaches to fuzzy clustering have been proposed. In the approaches, the objective functions are defined by using two different types of memberships. Fuzzy-Possibilistic c-Means (FPCM) algorithm [14] solves the noise sensitivity defect by introducing typicality values to the objective function of FCM. The algorithm simultaneously finds both membership values (relative typicalities) and typicality values (absolute typicalities) for each sample in the data set across all clusters in order to estimate good centroids alleviating the undesirable effects of outliers. For the task, the sum of the typicality values over all samples is required to be 1 (the row sum constraint).

Oh *et al.* [25] used two types of memberships in clustering of cooccurrence matrix in order to classify not only individuals but also categories. In order to group individuals and categories, which have high correlations each other, the algorithm tries to maximize the degree of aggregation, i.e., the total amount of products of qualitative variables and memberships for individuals and categories. To avoid trivial solutions, the total memberships of each cluster for categories are required to be 1. Umayahara *et al.* [26] also proposed several formulations of fuzzy clustering for categorical data based on multiset theory using two memberships and the row sum constraints.

In the next section, two types of memberships and the row sum constraints are introduced to linear fuzzy clustering in order to classify not only samples but also variables.

III. Local Principal Component Analysis Considering Responsibilities of Variables

A. Linear Fuzzy Clustering with Variable Selection

In order to eliminate unnecessary variables, the memberships of variables are introduced into the least squares criterion for Local PCA. This paper proposes two formulations based on two major fuzzification techniques. One is the standard approach adopted in the original FCM algorithm [3] that uses the fuzzifier of weighting exponent. The other is the entropy regularization approach [19] that is often associated with probabilistic mixture models [21].

Using two types of memberships, the objective function with the standard fuzzification approach is defined as follows:

$$\begin{aligned} L^s_{fcvvs} &= J_{(\theta_u,\theta_v)} \\ &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}^{\theta_u} \sum_{j=1}^{m} v_{cj}^{\theta_v} \Big(x_{ij} - \sum_{k=1}^{p} f_{cik} a_{cjk} - b_{cj} \Big)^2, \end{aligned} \quad (6)$$

where v_{cj} represents the degree of membership of the j-th variable to the c-th cluster. The weighting exponent $\theta_v (\theta_v > 1)$ plays a role for fuzzification of membership degrees of variables. If the j-th variable has no useful information for estimating the c-th prototypical linear variety, v_{cj} has small value and the j-th variable is ignored in calculation of clustering criteria in the c-th cluster.

On the other hand, using the entropy regularization approach, the objective function is given as

$$\begin{aligned} L^e_{fcvvs} &= J_{(1,1)} + \lambda_u \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci} \log u_{ci} \\ &\quad + \lambda_v \sum_{c=1}^{C}\sum_{j=1}^{m} v_{cj} \log v_{cj} \\ &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci} \sum_{j=1}^{m} v_{cj} \Big(x_{ij} - \sum_{k=1}^{p} f_{cik} a_{cjk} - b_{cj} \Big)^2 \\ &\quad + \lambda_u \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci} \log u_{ci} \\ &\quad + \lambda_v \sum_{c=1}^{C}\sum_{j=1}^{m} v_{cj} \log v_{cj}. \end{aligned} \quad (7)$$

The second entropy term plays a role for fuzzification of memberships of variables, and is controlled by the fuzzifier λ_v.

To obtain a unique solution, the objective functions are minimized under the constraints that

$$F_c^\top U_c^{\theta_u} F_c = I \quad ; \quad c = 1, \cdots, C, \quad (8)$$

$$F_c^\top U_c^{\theta_u} \mathbf{1}_n = \mathbf{0} \quad ; \quad c = 1, \cdots, C, \quad (9)$$

$$\sum_{c=1}^{C} u_{ci} = 1 \quad ; \quad i = 1, \cdots, n, \quad (10)$$

$$\sum_{j=1}^{m} v_{cj} = 1 \quad ; \quad c = 1, \cdots, C, \quad (11)$$

and $A_c^\top A_c$ is orthogonal. Here, the additional memberships represent the relative responsibilities of variables. So, the sum of v_{cj} is constrained to be 1 with respect to all m variables, but not with respect to all C clusters.

By the way, the additional membership of v_{cj} has a different role from that of the modified k-Means clustering proposed by Huang *et al.* [13] although the two objective functions have similar forms except for the clustering criteria. In [13], the memberships of variables were introduced in order to evaluate the importance of variables in clustering process. Then, the small memberships eliminate the effect of noisy variables. On the other hand, the proposed memberships v_{cj} work for evaluating the importance of variables in local linear model estimation in each cluster. So, the small memberships eliminate not only the effect of noisy variables in clustering process but also that of insignificant variables in linear model estimation.

The clustering result is derived by an iterative algorithm based on the alternate optimization technique. Here, the updating rules for parameters are first derived by using the standard fuzzification approach. To derive the optimal A_c, $\boldsymbol{b}_c$ and v_{cj}, Eq.(6) is rewritten as follows:

$$L^s_{fcvvs} = \sum_{c=1}^{C}\sum_{j=1}^{m} v_{cj}^{\theta_v} e_{cj}, \quad (12)$$

where

$$e_{cj} = (\boldsymbol{x}_j - F_c \tilde{\boldsymbol{a}}_{cj} - \mathbf{1}_n b_{cj})^\top U_c^{\theta_u} (\boldsymbol{x}_j - F_c \tilde{\boldsymbol{a}}_{cj} - \mathbf{1}_n b_{cj}), \quad (13)$$

$$A_c = (\tilde{\boldsymbol{a}}_{c1}, \cdots, \tilde{\boldsymbol{a}}_{cm})^\top.$$

From $\partial L^s_{fcvvs} / \partial \tilde{\boldsymbol{a}}_{cj} = \mathbf{0}$, and $\partial L^s_{fcvvs} / \partial b_{cj} = 0$, we have

$$\tilde{\boldsymbol{a}}_{cj} = (F_c^\top U_c^{\theta_u} F_c)^{-1} F_c^\top U_c^{\theta_u} (\boldsymbol{x}_j - \mathbf{1}_n b_{cj}), \quad (14)$$

and

$$b_{cj} = (\mathbf{1}_n^\top U_c^{\theta_u} \mathbf{1}_n)^{-1} \mathbf{1}_n^\top U_c^{\theta_u} (\boldsymbol{x}_j - F_c \tilde{\boldsymbol{a}}_{cj}). \quad (15)$$

The membership v_{cj} satisfying the probabilistic constraint of Eq.(11) is given as

$$\begin{aligned} v_{cj} &= \left\{ \sum_{l=1}^{m} \left(\frac{e_{cj}}{e_{cl}} \right)^{\frac{1}{\theta_v - 1}} \right\}^{-1} \\ &= \frac{(e_{cj})^{-\frac{1}{\theta_v - 1}}}{\sum_{l=1}^{m} (e_{cl})^{-\frac{1}{\theta_v - 1}}}. \end{aligned} \quad (16)$$

In the same way, we can derive the optimal F_c and u_{ci}. Eq.(6) is equivalent to

$$L^s_{fcvvs} = \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}^{\theta_u} d_{ci}, \quad (17)$$

where

$$d_{ci} = (\tilde{\boldsymbol{x}}_i - A_c \tilde{\boldsymbol{f}}_{ci} - \boldsymbol{b}_c)^\top \tilde{V}_c^{\theta_v} (\tilde{\boldsymbol{x}}_i - A_c \tilde{\boldsymbol{f}}_{ci} - \boldsymbol{b}_c). \quad (18)$$

$\partial L^s_{fcvvs} / \partial \tilde{\boldsymbol{f}}_{ci} = \mathbf{0}$ yields

$$\tilde{\boldsymbol{f}}_{ci} = (A_c^\top \tilde{V}_c^{\theta_v} A_c)^{-1} A_c^\top \tilde{V}_c^{\theta_v} (\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c), \quad (19)$$

where

$$\tilde{V}_c = \mathrm{diag}(v_{c1}, \cdots, v_{cm}). \quad (20)$$

The membership u_{ci} satisfying the probabilistic constraint of Eq.(10) is given as

$$u_{ci} = \left\{ \sum_{l=1}^{C} \left(\frac{d_{ci}}{d_{li}} \right)^{\frac{1}{\theta_u - 1}} \right\}^{-1} = \frac{(d_{ci})^{-\frac{1}{\theta_u - 1}}}{\sum_{l=1}^{C} (d_{li})^{-\frac{1}{\theta_u - 1}}}. \qquad (21)$$

Next, the updating rules with the entropy regularization approach are derived as follows: setting as $(\theta_u, \theta_v) = (1, 1)$, the optimal A_c, $\boldsymbol{b}_c$ and F_c are given as Eqs.(14), (15) and (19), respectively. In the same manner, from $\partial L^e_{fcvvs}/\partial v_{cj} = 0$ and $\partial L^e_{fcvvs}/\partial u_{ci} = 0$, we have

$$v_{cj} = \exp\left\{ -\frac{e_{cj}}{\lambda_v} - 1 \right\} \qquad (22)$$

and

$$u_{ci} = \exp\left\{ -\frac{d_{ci}}{\lambda_u} - 1 \right\}. \qquad (23)$$

The probabilistic constraints of Eqs.(11) and (10) normalize them as

$$v_{cj} = \frac{\exp\{-\frac{e_{cj}}{\lambda_v}\}}{\sum_{l=1}^{m} \exp\{-\frac{e_{cl}}{\lambda_v}\}} \qquad (24)$$

and

$$u_{ci} = \frac{\exp\{-\frac{d_{ci}}{\lambda_u}\}}{\sum_{l=1}^{C} \exp\{-\frac{d_{li}}{\lambda_u}\}}, \qquad (25)$$

respectively.

In this way, local subspaces are estimated ignoring unnecessary variables that have small memberships. However, when the number m of variables is large, the values will be very small because of the constraint of Eq.(11). So, it is often difficult to interpret the absolute responsibility of a variable from its responsibility value. The next subsection proposes to apply the graded possibilistic approach and make the additional memberships more useful for selecting variables.

B. Graded Possibilistic Approach to Variable Selection

The deficiency of the model proposed in the previous subsection comes from the fact that it imposes the same constraint with the conventional memberships on the variable selection parameters. In the mixed c-means clustering, Pal *et al.* [15] proposed to relax the constraint (row sum = 1) on the typicality values but retain the column constraint on the membership values. So, for deriving the absolute responsibilities of variables, the constraint is relaxed and the possibilistic constraint is applied in this subsection. In the possibilistic approach [24], the memberships can be regarded as the probability that an experimental outcome coincides with one of mutually independent events. However, it is possible that sets of events are neither mutually independent nor completely mutually exclusive. Then, Masulli and Rovetta [16], [17] proposed the graded possibilistic approach to clustering, in which soft transition of memberships from probabilistic to possibilistic constraint is performed by using the graded possibilistic constraint.

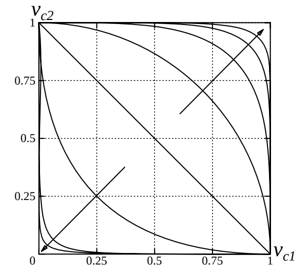


Fig. 2. Bounds of the feasible regions for v_{cj} for different values of α

In this subsection, the memberships of variables are estimated by using the graded possibilistic approach. Assume that a class of constraints is expressed by a unified formulation

$$\Psi = \sum_{j=1}^{m} v_{cj}^{[\xi]} - 1, \qquad (26)$$

where $[\xi]$ is an interval variable representing an arbitrary real number included in the range $[\underline{\xi}, \overline{\xi}]$, i.e., there must exist a scalar exponent $\xi^* \in [\underline{\xi}, \overline{\xi}]$ such that the equality $\Psi = 0$ holds. The constraint can be implemented by using a running parameter α. The extrema of the interval are written as a function of α, where

$$\underline{\xi} = \alpha, \qquad \overline{\xi} = \frac{1}{\alpha}, \qquad (27)$$

and $\alpha \in [0, 1]$. This formulation obviously includes two extreme cases. $\alpha = 1$ implies the probabilistic assumption because $[\xi] = [1, 1]$ yields $\sum_{j=1}^{m} v_{cj} = 1$. On the other hand, $\alpha = 0$ implies the possibilistic assumption because $[\xi] = [0, \infty]$ yields $\sum_{j=1}^{m} v_{cj}^0 \geq 1$ and $\sum_{j=1}^{m} v_{cj}^{\infty} \leq 1$. Then, the constraint with an interval is represented as a set of two inequalities.

$$\sum_{j=1}^{m} v_{cj}^{\alpha} \geq 1, \qquad \sum_{j=1}^{m} v_{cj}^{\frac{1}{\alpha}} \leq 1. \qquad (28)$$

Figure 2 depicts the bounds of the feasible regions for $m = 2$. The feasible value for v_{cj} must lie on the eye-shaped area, which becomes large in the direction of the arrows as the value of α decreases.

For implementation of the graded possibilistic clustering, a sample algorithm was introduced in [16], [17] considering the entropy regularization (or DA approach). In this section, the algorithm is redefined considering both of the standard fuzzification approach and the entropy regularization approach. Assume that we have the free membership function ϕ_{cj} that represents the absolute typicality of the j-th variable in the c-th cluster, and the membership value v_{cj} is given by the following normalization.

$$v_{cj} = \frac{\phi_{cj}}{\kappa_j}. \qquad (29)$$

Under the condition of $\phi_{cj} \in [0, 1]$, a reasonal choice of ϕ_{cj} for the entropy regularization approach is

$$\phi_{cj} = \exp\{-\frac{e_{cj}}{\lambda_v}\}. \qquad (30)$$

On the other hand, the standard fuzzification approach does not have such a free membership function. In this paper, the possibilistic membership [24] is used instead of $(e_{cj})^{-\frac{1}{\theta_v - 1}}$ of Eq.(16), then ϕ_{cj} is given as

$$\phi_{cj} = \frac{1}{1 + (e_{cj}/\eta_c)^{\frac{1}{\theta_v - 1}}}, \tag{31}$$

where η_c is a predefined constant. We can see that $e_{cj} = 0$ derives $\phi_{cj} = 1$, and ϕ_{cj} moves toward 0 as $e_{cj} \to \infty$.

In the graded possibilistic approach, κ_j takes one of the following three values.

$$\kappa_j = \left(\sum_{l=1}^{m} \phi_{cl}^{\frac{1}{\alpha}} \right)^{\alpha} \qquad \text{if} \sum_{l=1}^{m} \phi_{cl}^{\frac{1}{\alpha}} > 1, \tag{32}$$

$$\kappa_j = \left(\sum_{l=1}^{m} \phi_{cl}^{\alpha} \right)^{\frac{1}{\alpha}} \qquad \text{if} \sum_{l=1}^{m} \phi_{cl}^{\alpha} < 1, \tag{33}$$

$$\kappa_j = 1 \qquad \text{else.} \tag{34}$$

Eq.(32) transforms memberships that are above the eye-shaped area onto the upper boundary ($\sum_{l=1}^{m} v_{cl}^{\frac{1}{\alpha}} = 1$) while Eq.(33) transforms memberships under the eye-shaped area onto the lower boundary ($\sum_{l=1}^{m} v_{cl}^{\alpha} = 1$). Here, $\sum_{l=1}^{m} \phi_{cl}^{\frac{1}{\alpha}} > 1$ and $\sum_{l=1}^{m} \phi_{cl}^{\alpha} < 1$ are mutually exclusive because $\phi_{cl}^{\alpha} > \phi_{cl}^{\frac{1}{\alpha}}$ derives

$$\sum_{l=1}^{m} \phi_{cl}^{\alpha} < 1 \Rightarrow \sum_{l=1}^{m} \phi_{cl}^{\frac{1}{\alpha}} \leq 1, \tag{35}$$

and

$$\sum_{l=1}^{m} \phi_{cl}^{\frac{1}{\alpha}} > 1 \Rightarrow \sum_{l=1}^{m} \phi_{cl}^{\alpha} \geq 1. \tag{36}$$

When $\alpha = 1$, Eqs.(32) and (33) derive $\kappa_j = \sum_{l=1}^{m} \phi_{cl}$, and memberships v_{cj}'s are reduced to the probabilistic ones ($\sum_{j=1}^{m} v_{cj} = 1$). On the other hand, $\alpha = 0$ provides the possibilistic membership assignment because all of κ_j's are given by Eq.(34). Then, the value of α should be gradually decreased from 1 to 0.

The proposed algorithm can be written as follows:

Algorithm: Fuzzy c-Varieties with Variable Selection Using Graded Possibilistic Approach

Step 1 Initialize $U_c, \tilde{V}_c, A_c, \boldsymbol{b}_c, F_c$ in each cluster and normalize them so that they satisfy the constraints Eqs.(8)-(11) and $A_c^{\top} A_c$ is orthogonal. Set the running parameter as $\alpha = 1$. Choose termination threshold ε and the step of running parameter $\Delta\alpha$.

Step 2 Update A_c's using Eq.(14) and transform them so that each $A_c^{\top} A_c$ is orthogonal.

Step 3 Update F_c's using Eq.(19) and normalize them so that they satisfy the constraints Eqs.(8) and (9).

Step 4 Update $\boldsymbol{b}_c$'s using Eq.(15).

Step 5 Update $\tilde{V}_c$'s using Eq.(29).

Step 6 Update U_c's using Eq.(21) (or Eq.(25)).

Step 7 If

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

then go to Step 8. Otherwise, return to Step 2.

Step 8 If $\alpha = 0$, then stop. Otherwise, set $\alpha = \alpha - \Delta\alpha$ and return to Step 2.

When we use the entropy regularization approach, the above algorithm with $\alpha = 1$ is equivalent to that of probabilistic constraint model. So, the algorithm exactly performs the soft transition from the probabilistic clustering model to the possibilistic one. On the other hand, when we use the standard fuzzification approach, the algorithm is not equivalent to that of probabilistic one even if $\alpha = 1$. Then, we should use the result of the probabilistic clustering model in the initialization step. Furthermore, we must also choose the additional parameter η_c carefully.

C. Comparison with Robust FCV

Although the Robust FCV algorithm [12] is a technique for handling intra-sample outliers, it can also be extended to variable selection in linear fuzzy clustering. In this subsection, the close relationship between the proposed method and the robust M-estimation is discussed. In the proposed variable selection model, the goal is to estimate linear models ignoring unnecessary variables. In order to derive robust models that are free from the influences of outliers, the objective functions of least squares techniques have been enhanced to the robust measures using robust ρ-functions [22]. Based on robust M-estimation, the energy function to be minimized is defined as follows:

$$L'_{rfcv} = \sum_{c=1}^{C} \sum_{j=1}^{m} \rho(e_{cj}^{1/2}) + \lambda \sum_{c=1}^{C} \sum_{i=1}^{n} u_{ci} \log u_{ci}, \tag{37}$$

where $\rho(\cdot)$ is a class of robust ρ-functions, and e_{cj} is the previous residual for the j-th variable given by Eq.(13). To solve the minimization problem, both the iteratively reweighted least-squares (IRLS) technique [22] and the gradient descent method with a local quadratic approximation are used. For the Geman-McClure ρ function [27]

$$\rho(x) = \frac{x^2}{x^2 + \sigma^2}, \tag{38}$$

the weight corresponding to the j-th variable w_{cj} is given by

$$w_{cj} = \frac{\psi(e_{cj}, \sigma)}{e_{cj}}, \tag{39}$$

where

$$\psi(e_{cj}, \sigma) = \frac{\partial \rho(e_{cj}^{1/2})}{\partial e_{cj}} = \frac{2 e_{cij} \sigma^2}{(e_{cij}^2 + \sigma^2)^2}, \tag{40}$$

and σ is a scale parameter that controls the convexity of the robust function.

Using the responsibility weight of the j-th variable in the c-th cluster, the objective function of the Robust FCV algorithm is redefined as

$$\begin{aligned} L''_{rfcv} &= \sum_{c=1}^{C}\sum_{j=1}^{m} w_{cj}e_{cj} + \lambda \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci} \\ &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\sum_{j=1}^{m} w_{cj}\Big(x_{ij} - \sum_{k=1}^{p} f_{cik}a_{cjk} - b_{cj}\Big)^2 \\ &\quad +\lambda \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci}. \end{aligned} \tag{41}$$

Here, let ϕ be the vector of parameters. The first derivative with fixed w_{cj} is similar to that of L'_{rfcv} as follows:

$$\begin{aligned} \frac{\partial L''_{rfcv}}{\partial \phi} &= w_{cj}\frac{\partial e_{cj}}{\partial \phi} \\ &= \frac{\sigma^2}{(e_{cj}+\sigma^2)^2}\frac{\partial e_{cj}}{\partial \phi} \cong \frac{\partial L'_{rfcv}}{\partial \phi}. \end{aligned} \tag{42}$$

Therefore, minimization of the modified objective function of Eq.(41) approximately achieves the optimization of Eq.(37), and the proposed hybrid mechanism for variable selection in linear fuzzy clustering, which uses two types of memberships, has close relationship with the robust M-estimation that uses the responsibility weights of variables based on the IRLS technique.

By the way, it should be noted that the robust model of Eq.(37) uses the weighted sum of squared errors with respect to the j-th variable as the measure to be minimized and robustifies it by using robust ρ-function. In this sense, the responsibility weight w_{cj} plays a role for robustification using the robust M-estimation technique while the membership weight u_{ci} fuzzifies the distance measure. Tuning of the responsibility weight, however, is not performed by minimization of a single objective function. On the other hand, in the proposed linear fuzzy clustering algorithm, the robustification mechanism is performed by using two different memberships and the memberships are calculated by the minimization of a single objective function.

D. Selection of Fuzzifier for Memberships of Variables

In the standard fuzzification method, the fuzzifier is often set as $\theta = 2.0$ or sometime as $\theta = 1.5$ in view of clear classification. On the other hand, the entropy regularization approach does not have such general choice of the fuzzifier because the degree of fuzziness is depends not only on the value of fuzzifier but also on the scale of variables. Although several techniques have been used for deriving the optimal solutions for some regularization problems [28], we cannot derive the optimal fuzzifier for the entropy regularizer using such techniques because the entropy regularization is not exactly a technique for regularizing ill-posed problems but for fuzzification of memberships in the clustering context. So, the degree of fuzziness should be decided based on the standpoint of analyst in the same way as the standard fuzzification technique. This subsection proposes an approach to automatic selection of the fuzzifier for the memberships of variables so as to reflect the view point of analyst.

It has been shown that the FCM algorithm with entropy regularization method has close relation with the EM algorithm for Gaussian Mixture Models (GMMs) and the iterative algorithm is equivalent to the EM algorithm for GMMs in the case where the unknown parameters of Gaussian components are only the mean vectors [21]. In the model, fuzzifier λ is identified with the double variance of the Gaussian component density functions. The regularization by K-L information [21] is an extended version of the entropy regularization that is comparable with GMMs with full parameters. In the technique, the fuzzifier λ is used for generalizing GMMs considering the analyst's view point.

The memberships of variables proposed in this paper are also comparable with a Gaussian model. Although the goal of clustering is not the estimation of a probability density function, we can consider a corresponding probabilistic model based on the algorithmic similarity. Assume that the previous residuals for variables $e_{cj}^{1/2}, j = 1, \cdots, m$ are drawn from a probabilistic distribution and we have a Gaussian density function $p_c(\cdot)$ with fixed σ_{vc}^2 in the c-th cluster.

$$p_c(e_{cj}^{1/2}) = \frac{1}{\sqrt{2\pi}\sigma_{vc}} \exp\left(\frac{-e_{cj}}{2\sigma_{vc}^2}\right), \tag{43}$$

where e_{cj} is the squared deviation for the j-th object. The free membership of Eq.(30) is the normalized version of $p_c(e_{cj}^{1/2})$ such that $p_c(0) = 1$, and the fuzzifier for the memberships is equivalent to the double variance of the deviation ($\lambda_v = 2\sigma_{vc}^2$).

Here, we can derive a possible approach to automatic adjustment of the fuzzifier based on the sample variances considering the analyst's view point. In order to adjust the fuzzifier, the fuzzifier is decomposed as $\lambda_v = \Lambda_v \sigma_{vc}^2$ where λ_v can be automatically adjusted in each cluster by updating σ_{vc}^2 while Λ_v is a constant for tuning the degree of fuzziness.

$$\sigma_{vc}^2 = \frac{1}{m}\sum_{j=1}^{m} e_{cj}. \tag{44}$$

Then, if $\Lambda_v = 2$, the possibilistic assignment of memberships of variables corresponds to the Gaussian model.

Here, $\Lambda_v = 2$ is a natural choice from the view point of probabilistic clustering [29] and Λ_v should be set smaller than 2 for clear classification.

IV. Numerical Experiments

A. Analysis of an Artificial Data Set

A numerical experiment was performed using an artificial data set. Table I shows the coordinates of the samples. Samples 1-12 form the first group, in which x_1, x_2 and x_3 are linearly related, i.e., samples are distributed forming a line in the 3-D space. However, x_4 and x_5 are random variables. So, x_1, x_2 and x_3 should be selected in the group and we can capture the local linear structure by eliminating x_4 and x_5. On the other hand, samples 13-24 form the second group, in which x_2, x_3 and x_4 are linearly related, but x_1 and x_5 are random variables. In this way, the local structures must be captured by classifying not only samples but also variables.

First, applying the proposed algorithm with the standard fuzzification method, the samples were partitioned into two

TABLE I
ARTIFICIAL DATA SET

sample	x_1	x_2	x_3	x_4	x_5
1	0.000	0.000	0.250	**0.143**	**0.365**
2	0.091	0.091	0.295	**0.560**	**0.605**
3	0.182	0.182	0.341	**0.637**	**0.001**
4	0.273	0.273	0.386	**0.529**	**0.557**
5	0.364	0.364	0.432	**0.949**	**0.195**
6	0.455	0.455	0.477	**0.645**	**0.206**
7	0.545	0.545	0.523	**0.598**	**0.026**
8	0.636	0.636	0.568	**0.616**	**0.729**
9	0.727	0.727	0.614	**0.004**	**0.407**
10	0.818	0.818	0.659	**0.255**	**0.641**
11	0.909	0.909	0.705	**0.088**	**0.244**
12	1.000	1.000	0.750	**0.589**	**0.213**
13	**0.199**	0.750	0.250	0.000	**0.321**
14	**0.411**	0.705	0.295	0.091	**0.167**
15	**0.365**	0.659	0.341	0.182	**0.419**
16	**0.950**	0.614	0.386	0.273	**0.109**
17	**0.581**	0.568	0.432	0.364	**0.561**
18	**0.323**	0.523	0.477	0.455	**0.127**
19	**0.899**	0.477	0.523	0.545	**0.349**
20	**0.399**	0.432	0.568	0.636	**0.100**
21	**0.249**	0.386	0.614	0.727	**0.682**
22	**0.214**	0.341	0.659	0.818	**0.714**
23	**0.838**	0.295	0.705	0.909	**0.605**
24	**0.166**	0.250	0.750	1.000	**0.244**

clusters using the probabilistic constraint. The model parameters were set as $\theta_u = 2.0, \theta_v = 2.0, p = 1$ and $\alpha = 1$. In the sense of maximum membership, the first cluster included samples 1-12, while the second cluster included the remaining samples. The left columns of Table II show the derived memberships of variables and local principal component vectors. x_4 and x_5 were eliminated in the first cluster and $\boldsymbol{a}_1$ revealed the relationship among x_1, x_2 and x_3. On the other hand, in the second cluster, small memberships were assigned to x_1 and x_5, and $\boldsymbol{a}_2$ represented the local structure of the second group (samples 13-24). A similar clustering result can be also derived by using the entropy regularization method. The result is shown in the left columns of Table III. The model parameters were set as $\lambda_u = 0.01, \lambda_v = 0.5, p = 1$ and $\alpha = 1$.

The clustering results indicate that the proposed membership v_{cj} is useful for evaluating the typicality of the variable in local linear model estimation where x_1 and x_4 are significant only in the first and second cluster, respectively. Additionally, the typicality values also play a role for rejecting the influences of noise variable (x_5) because x_5 belongs to neither of two clusters. In this way, the row sum constraints of Eq.(11) give the memberships a different meaning from the conventional column constraints that forces each samples to belong at least one cluster. However, the memberships represent only the relative responsibilities and cannot be used for determining the absolute responsibilities.

Then, the graded possibilistic approach was applied to the data set. The step of the running parameter was set as $\Delta\alpha = 0.1$. The right columns of Tables II and III show the derived memberships of variables and local principal component vectors. Because the responsibility of each variable is represented by the possibilistic partition, the membership values make it easy to select the variables to be considered. In this way, the proposed algorithm is useful for variable selection in linear fuzzy clustering.

By the way, in the standard fuzzification method, we must choose additional parameter η_c. In this experiment, the result of Table II was given with $\eta_c = 0.05$. However, the clustering result severely depended on the value of η_c. So, a careful choice is required in order to derive meaningful membership values using the graded possibilistic approach.

In the entropy regularization method, we can evaluate the degree of fuzziness considering the connection with the gaussian mixture models [21]. When we use the probabilistic constraint, the fuzzifier corresponds to the variance of the gaussian distribution, and the clustering model is equivalent to the gaussian mixture models if the fuzzifier is equal to the double variance of the corresponding component distribution. In the proposed clustering model, the generalized variance of the c-th cluster is calculated as

$$\sigma_{uc}^2 = \frac{\sum_{i=1}^n u_{ci} d_{ci}}{\sum_{i=1}^n u_{ci}}, \tag{45}$$

and the values were given as $(\sigma_{u1}^2, \sigma_{u2}^2) = (0.01, 0.01)$ in this experiment. So, it can be said that the derived fuzzy partition with $\lambda_u = 0.1$ is slightly fuzzy compared with the gaussian mixture models. On the other hand, in the possibilistic model, the membership function corresponds to single gaussian model and the fuzzifier plays a role of the variance of the distribution. For memberships of variables, the variance of the c-th cluster were $(\sigma_{v1}^2, \sigma_{v2}^2) = (0.15, 0.14)$. So, the value of $\lambda_v = 0.05$ was a strict choice, in which the degree of importance is more emphasized than the gaussian distribution. In this way, the clustering model with the entropy regularization is comparable with its corresponding probabilistic model.

B. Analysis of Wine Data

Next, the proposed method was applied to Wine data [30], which consists of 178 instances with 13 numerical attributes and 1 class label. In this experiment, the entropy regularization method was used for the fuzzification of memberships. Each instance belongs to one of three classes. This data set is often used for pattern recognition tasks. However, in this experiment, it is used for an unsupervised classification task. So, the goal of the analysis is to reveal the local structure of the data set using only 13 attributes (without class label). Before application of clustering algorithms, the data set was normalized so that each attribute has zero mean and unit variance.

First, the data set was partitioned into two clusters using the conventional FCV algorithm. The left side of Table IV shows the average of memberships in each cluster. In the sense of maximum membership, all instances of the first class were assigned into the first cluster while the second cluster included the third class. The instances of the second class were shared by the two clusters. Then, the first (second) cluster reveals the relationship between the first and second classes (the second and third classes), i.e., the data set can be summarized by two linear structures, and the second class is located in the intersection of the two lines. Table V shows the cluster centers and the fuzzy factor loadings (FFL) of the component scores.

TABLE II

MEMBERSHIPS OF VARIABLES AND LOCAL PRINCIPAL COMPONENT VECTORS WITH STANDARD FUZZIFICATION (ARTIFICIAL DATA SET)

variable	probabilistic constraint				possibilistic constraint			
	v_{cj}		$\boldsymbol{a}_c$		v_{cj}		$\boldsymbol{a}_c$	
	$c=1$	$c=2$	$c=1$	$c=2$	$c=1$	$c=2$	$c=1$	$c=2$
x_1	0.318	0.018	1.083	0.036	0.833	0.055	1.083	0.036
x_2	0.318	0.316	1.083	0.539	0.833	0.833	1.083	0.539
x_3	0.318	0.316	0.541	-0.539	0.833	0.833	0.541	-0.539
x_4	0.020	0.316	-0.236	-1.079	0.059	0.833	-0.236	-1.079
x_5	0.026	0.035	0.007	-0.291	0.076	0.099	0.007	-0.291

TABLE III

MEMBERSHIPS OF VARIABLES AND LOCAL PRINCIPAL COMPONENT VECTORS WITH ENTROPY REGULARIZATION (ARTIFICIAL DATA SET)

variable	probabilistic constraint				possibilistic constraint			
	v_{cj}		$\boldsymbol{a}_c$		v_{cj}		$\boldsymbol{a}_c$	
	$c=1$	$c=2$	$c=1$	$c=2$	$c=1$	$c=2$	$c=1$	$c=2$
x_1	0.283	0.056	1.103	0.044	0.955	0.192	1.103	0.044
x_2	0.287	0.280	1.073	0.540	0.968	0.966	1.073	0.540
x_3	0.279	0.272	0.522	-0.505	0.941	0.937	0.522	-0.505
x_4	0.069	0.283	-0.242	-1.104	0.233	0.976	-0.242	-1.104
x_5	0.082	0.109	-0.001	-0.289	0.278	0.376	-0.001	-0.289

FFL is the correlation coefficient between the component score and the original attribute [6] and can be used for selecting meaningful attributes whose FFL values are shown in bold. For example, variables x_4 and x_5 are meaningful only in the first cluster while x_2, x_3, x_{11} and x_{12} are selected in second cluster. However, it is not so easy to evaluate the responsibility of each variable for local model estimation, and the local principal components might be influenced by meaningless variables, i.e., we cannot evaluate the amount of information kept in the local models using only the fuzzy factor loadings from the view point of prediction. Furthermore, in order to keep in meaningful information as well as possible, the local models should be estimated by eliminating the influences of insignificant variables.

Second, the proposed method was applied in order to partition the data set into two clusters selecting variables in each cluster. The model parameters were set as $\lambda_u = 0.05, \lambda_v = 50.0$, $p = 1$ and $\Delta\alpha = 0.1$. As the result, the variances of Eqs.(45) and (44) were given as $(\sigma_{u1}^2, \sigma_{u2}^2) = (0.40, 0.33)$ and $(\sigma_{v1}^2, \sigma_{v2}^2) = (41.66, 45.05)$, respectively. Then, the parameter setting is very crisp for sample partition and slightly strict for variable partition compared with gaussian distribution. The right side of Table IV shows the average of memberships in each cluster. In the sense of maximum membership, the membership assignments are similar to that of the FCV clustering. Tables VI shows the cluster centers, the memberships of variables and the fuzzy factor loadings of the component scores derived with possibilistic constraint. The variables indicated by bold symbols were assigned large memberships. Here, the tables show the following two notable features. (i) the variables whose FFL values had large magnitude in Table V were assigned large memberships, i.e., the variables that has high correlation with the local latent variable are regarded as important. So, we can select variables that are mutually dependent, e.g., x_1, x_6, x_7, x_{10} and x_{13} can be used for constructing association rules in the first cluster. (ii) the variables corresponding to the coordinate with wide gaps in cluster centers, such as x_2 and x_{12} in the first cluster and x_{13} in the second cluster, were assigned large memberships, i.e., the variables that are mainly used for data partitioning are regarded as important. The first cluster consists of the samples whose x_2 (x_{12}) is small (large), while the second cluster is characterized by samples whose x_{13} is small. So, we can also select variables that are useful for understanding the cluster structure. Furthermore, from the view point of prediction, the derived local linear models have a different feature from that of FCV because the models are biased by the additional membership so that the models keep in meaningful information as well as possible. Table VII shows the comparison of the membership-weighted mean squared error (MSE) of each variable. The variables having large memberships are shown in bold. Because the large memberships try to minimize the errors for the meaningful (i.e., predictable) variables, the corresponding MSE became smaller than the original model given by FCV in most cases. Then, it can be said that the local linear models derived by the proposed method are more useful for revealing the mutual structure of the predictable variables by giving up representation of unpredictable variables. In this way, the proposed method emphasizes the effects of meaningful variables in local linear modeling, and the memberships of variables are used for selecting meaningful variables.

Third, the automatic adjustment mechanism was applied in order to compare the results derived with several parameter sets. In this experiment, the memberships of samples u_{ci}'s were fixed as the previous result and only v_{cj}'s were updated in the soft transition process. Changing the value of Λ_v, the memberships of variables were given as Table VIII. The table shows that the classification result becomes clear as Λ_v becomes small, i.e., the deviations between the memberships of meaningful and insignificant variables gradually grow as Λ_v becomes small while $\Lambda_v \to \infty$ corresponds to the standard FCV clustering where v_{cj}'s are all 1. However, in the result of $\Lambda_v = 0.5$, some memberships became unusually larger

TABLE IV
AVERAGE OF MEMBERSHIPS IN EACH CLUSTER (WINE DATA)

class label	FCV		Proposed method	
	$c=1$	$c=2$	$c=1$	$c=2$
1	0.978	0.022	0.973	0.027
2	0.446	0.554	0.559	0.441
3	0.034	0.966	0.021	0.979

TABLE V
CLUSTER CENTERS AND FUZZY FACTOR LOADINGS OF COMPONENT SCORES DERIVED BY FCV ALGORITHM (WINE DATA)

variable	$\boldsymbol{b}_c$		FFL	
	$c=1$	$c=2$	$c=1$	$c=2$
x_1	0.241	-0.252	**0.871**	**0.619**
x_2	-0.298	0.312	-0.048	**0.698**
x_3	0.275	-0.288	-0.049	**0.590**
x_4	-0.239	0.250	**-0.698**	0.451
x_5	0.178	-0.187	**0.514**	0.163
x_6	0.512	-0.536	**0.785**	**-0.667**
x_7	0.608	-0.637	**0.768**	**-0.817**
x_8	-0.224	0.234	**-0.661**	**0.603**
x_9	0.410	-0.430	0.442	-0.484
x_{10}	-0.149	0.156	**0.755**	**0.776**
x_{11}	0.410	-0.429	0.019	**-0.835**
x_{12}	0.547	-0.573	0.414	**-0.855**
x_{13}	0.496	-0.519	**0.872**	0.334

than those with $\Lambda_v = 1.0$. So, we can see that the result of $\Lambda_v = 0.5$ corresponds to one of inappropriate local solutions and $\Lambda_v = 1.0$ should be used for clear classification while $\Lambda_v = 2.0$ gives a general result from the view point of probabilistic classification. In this sense, the previous result where the degree of fuzziness is nearly equal to $\Lambda_v = 1.0$ is a fair solution for clear classification. In this way, the automatic adjustment mechanism is useful for evaluating the degree of fuzziness.

TABLE VI
CLUSTER CENTERS, MEMBERSHIPS OF VARIABLES AND FUZZY FACTOR LOADINGS OF COMPONENT SCORES DERIVED BY PROPOSED METHOD USING GRADED POSSIBILISTIC APPROACH (WINE DATA)

variable	$\boldsymbol{b}_c$		FFL		v_{cj}	
	$c=1$	$c=2$	$c=1$	$c=2$	$c=1$	$c=2$
x_1	**0.199**	**-0.245**	**0.896**	**0.683**	**0.633**	**0.575**
x_2	**-0.474**	0.583	**0.213**	0.377	**0.572**	0.184
x_3	-0.010	0.012	0.285	0.165	0.111	0.326
x_4	-0.346	0.425	-0.469	0.101	0.209	0.367
x_5	0.120	-0.147	0.433	0.095	0.174	0.270
x_6	**0.387**	**-0.475**	**0.875**	**-0.857**	**0.716**	**0.678**
x_7	**0.513**	**-0.630**	**0.867**	**-0.919**	**0.770**	**0.813**
x_8	-0.283	0.348	-0.490	0.476	0.350	0.247
x_9	0.295	-0.363	0.534	-0.668	0.333	0.407
x_{10}	**-0.183**	0.225	**0.811**	0.727	**0.665**	0.354
x_{11}	0.561	**-0.689**	-0.179	**-0.713**	0.439	**0.524**
x_{12}	**0.511**	**-0.628**	**0.454**	**-0.921**	**0.510**	**0.791**
x_{13}	**0.428**	**-0.526**	**0.873**	**0.357**	**0.572**	**0.716**

TABLE VII
COMPARISON OF MEAN SQUARED ERRORS (WINE DATA)

variable	FCV		proposed method			
	MSE		MSE		v_{cj}	
	$c=1$	$c=2$	$c=1$	$c=2$	$c=1$	$c=2$
x_1	**0.291**	**0.398**	**0.233**	**0.346**	**0.634**	**0.575**
x_2	**0.548**	0.650	**0.285**	1.058	**0.572**	0.184
x_3	0.763	0.697	1.119	0.700	0.112	0.326
x_4	0.593	0.560	0.797	0.627	0.210	0.367
x_5	0.736	0.895	0.893	0.819	0.174	0.269
x_6	**0.248**	**0.444**	**0.170**	**0.245**	**0.717**	**0.676**
x_7	**0.199**	**0.245**	**0.133**	**0.129**	**0.771**	**0.813**
x_8	0.434	0.714	0.533	0.877	0.352	0.246
x_9	0.603	0.681	0.560	0.563	0.334	0.406
x_{10}	**0.272**	0.528	**0.208**	0.649	**0.666**	0.354
x_{11}	0.534	**0.337**	0.418	**0.404**	0.441	**0.524**
x_{12}	**0.370**	**0.250**	**0.343**	**0.147**	**0.511**	**0.791**
x_{13}	**0.289**	**0.221**	**0.285**	**0.209**	**0.572**	**0.715**

V. CONCLUSION

This paper proposed a new approach to linear fuzzy clustering that selects variables considering the responsibility of variables. The responsibility of each variable for local model estimation is represented by the additional memberships, which are estimated by the graded possibilistic approach to clustering. The additional memberships are used not only for evaluating the mutual dependencies of variables but also for interpreting cluster structure, and are calculated using the reconstruction errors in the lower rank estimation. So, the new measure is also responsible for estimating the prediction ability in reduced-rank prediction, and the proposed algorithm tries to keep the information of meaningful (i.e., predictable) variables as well as possible. Then, the proposed method can be applied to missing value estimation problems, such as collaborative filtering [10], [12], in order to improve the prediction abilities for meaningful variables.

The close relationship with the robust M-estimation indicates that the memberships of variables also play a role of the responsibility weights that approximately achieves the optimization of robust ρ-function, in which the weighted sum of squared errors with respect to the variables are used as a measure of fit.

The flexibility of the model, however, forces the analysts to determine the additional parameters such as the degree of fuzziness and the step of running parameter. Usually, the measure for model selection depends on applications and there exists no general measure for clustering. In this paper, the degree of fuzziness was evaluated considering the connection between the entropy regularization method and the gaussian mixture models [21]. While we have several criteria for evaluating the FCM partition, most of them can not be applied to linear fuzzy clustering without modification because they do not consider the model complexity, i.e., the dimensionality of local subspaces. It is expected to establish a general measure for both of the model complexity and the degree of fuzziness.

Another possible future work is to develop the graded possibilistic mechanisms using other fuzzification techniques [21], [31], [32]. In this paper, two types of free membership functions were proposed, and the entropy regularization approach made it possible to enhance the probabilistic partition to the possibilistic one directly while the standard fuzzification method needed to redefine the free membership function. We can also define other graded possibilistic models by using other fuzzification method, and it is possible to apply different fuzzification method to the two memberships. However, we

TABLE VIII

COMPARISON OF MEMBERSHIPS OF VARIABLES DERIVED WITH SEVERAL PARAMETER SETS (WINE DATA)

variable	$\Lambda_v = 0.5$		$\Lambda_v = 1.0$		$\Lambda_v = 2.0$		$\Lambda_v = 3.0$		$\Lambda_v = 5.0$		$\Lambda_v = 100.0$	
	$c = 1$	$c = 2$	$c = 1$	$c = 2$	$c = 1$	$c = 2$	$c = 1$	$c = 2$	$c = 1$	$c = 2$	$c = 1$	$c = 2$
x_1	0.234	0.213	**0.604**	**0.509**	**0.769**	**0.725**	**0.835**	**0.809**	**0.895**	**0.882**	**0.994**	**0.994**
x_2	0.299	0.046	**0.538**	0.130	**0.733**	0.368	**0.813**	**0.518**	**0.883**	**0.678**	**0.994**	**0.981**
x_3	0.009	**0.809**	0.088	0.261	0.294	**0.512**	0.442	**0.641**	**0.614**	**0.767**	**0.976**	**0.987**
x_4	0.028	**0.662**	0.176	0.301	0.424	**0.548**	**0.566**	**0.670**	**0.711**	**0.787**	**0.983**	**0.988**
x_5	0.020	0.121	0.142	0.209	0.389	0.452	**0.539**	**0.588**	**0.696**	**0.726**	**0.983**	**0.984**
x_6	**0.681**	0.088	**0.696**	**0.643**	**0.821**	**0.767**	**0.873**	**0.830**	**0.920**	**0.890**	**0.996**	**0.994**
x_7	**0.706**	0.110	**0.752**	**0.795**	**0.858**	**0.865**	**0.900**	**0.901**	**0.938**	**0.936**	**0.997**	**0.996**
x_8	0.097	0.053	0.312	0.185	**0.563**	0.436	**0.683**	**0.579**	**0.795**	**0.724**	**0.989**	**0.984**
x_9	0.098	0.066	0.295	0.346	**0.547**	**0.571**	**0.671**	**0.684**	**0.788**	**0.794**	**0.988**	**0.988**
x_{10}	0.405	0.042	**0.639**	0.273	**0.790**	**0.562**	**0.852**	**0.688**	**0.907**	**0.802**	**0.995**	**0.989**
x_{11}	0.171	0.132	0.403	0.452	**0.633**	**0.688**	**0.737**	**0.782**	**0.832**	**0.863**	**0.991**	**0.993**
x_{12}	0.245	0.080	0.475	**0.764**	**0.687**	**0.851**	**0.778**	**0.890**	**0.860**	**0.928**	**0.992**	**0.996**
x_{13}	0.178	**0.531**	**0.535**	**0.671**	**0.737**	**0.816**	**0.815**	**0.873**	**0.884**	**0.921**	**0.994**	**0.996**

cannot always define the free membership function for every fuzzification method. The comparative study would be helpful for users in selection of the clustering models.

ACKNOWLEDGMENT

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

REFERENCES

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

[2] 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.

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

[4] 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.

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

[6] Y. Yabuuchi and J. Watada, "Fuzzy principal component analysis and its application," *Biomedical Fuzzy and Human Sciences*, vol.3, pp.83-92, 1997.

[7] I. T. Jolliffe, "Discarding variables in a principal component analysis. I. Artificial data," *Appl. Statist.*, vol.21, pp.160-173, 1972.

[8] Y. Tanaka and Y. Mori, "Principal component analysis based on a subset of variables: variable selection and sensitivity analysis," *American Journal of Mathematics and Management Sciences*, vol.17, no.1&2, pp.61-89, 1997.

[9] VASpca (VAriable Selection in Principal Component Analysis) Web Page; http://mo161.soci.ous.ac.jp/vaspca/indexE.html

[10] K. Honda, N. Sugiura, H. Ichihashi and S. Araki, "Collaborative filtering using principal component analysis and fuzzy clustering," *Web Intelligence: Research and Development, Lecture Notes in Artificial Intelligence 2198*, Springer, pp. 394-402, 2001.

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

[12] K. Honda and H. Ichihashi, "Component-wise robust linear fuzzy clustering for collaborative filtering," *International Journal of Approximate Reasoning*, vol.37, no.2, pp.127-144, 2004.

[13] J. Z. Huang, M. K. Ng, H. Rong and Z. Li, "Automated variable weighting in k-means type clustering," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.27, no.5, pp.657-668, 2005.

[14] N. R. Pal, K. Pal and J. C. Bezdek, "A mixed c-means clustering model," *Proc. of 1997 IEEE Int. Conf. on Fuzzy Systems*, pp.11-21, 1997.

[15] N. R. Pal, K. Pal, J. M. Keller and J. C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE Transactions on Fuzzy Systems*, vol.13, no.4, pp.508-516, 2005.

[16] F. Masulli and S. Rovetta, "The graded possibilistic clustering model," *IJCNN 2003 Conference Proceedings*, pp.791-796, 2003.

[17] F. Masulli and S. Rovetta, "Soft transition from probabilistic to possibilistic fuzzy clustering," *IEEE Transactions on Fuzzy Systems*, (in press).

[18] P. Whittle, "On principal components and least square methods of factor analysis," *Skand. Akt.*, vol.36, pp.223-239, 1952.

[19] 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.

[20] K. Rose, E. Gurewitz, and G. Fox, "A deterministic annealing approach to clustering," *Pattern Recognition Letters*, vol. 11, pp. 589-594, 1990.

[21] K. Honda and H. Ichihashi, "Regularized linear fuzzy clustering and probabilistic PCA mixture models," *IEEE Trans. on Fuzzy Systems*, vol. 13, no. 4, pp. 508-516, 2005.

[22] P. W. Holland and R. E. Welsch, "Robust regression using iteratively reweighted least-squares," *Communications in Statistics*, vol.A6, no.9, pp.813-827, 1977.

[23] R. N. Davé and R. Krishnapuram, "Robust clustering methods: a unified view," *IEEE Trans. on Fuzzy Systems*, vol.5, pp.270-293, 1997.

[24] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Trans. on Fuzzy Systems*, vol.1, pp.98-110, 1993.

[25] C.-H. Oh, K. Honda and H. Ichihashi, "Fuzzy clustering for categorical multivariate data," *Proc. of Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, pp.2154-2159, 2001.

[26] K. Umayahara, S. Miyamoto and Y. Nakamori, "Formulations of fuzzy clustering for categorical data," *International Journal of Innovative Computing, Information and Control*, vol. 1, no.1, pp.83-94, 2005.

[27] S. Geman, D. E. McClure, "Statistical methods for tomographic image reconstruction," *Bulletin of International Statistical Institute*, vol.LII-4, pp.5-21, 1987.

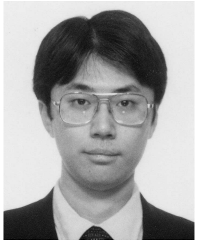
[28] D. L. Pham, "Spatial models for fuzzy clustering," *Computer Vision and Image Understanding*, vol.84, pp.285-297, 2001.

[29] R. J. Hathaway, "Another interpretation of the EM algorithm for mixture distributions," *Statistics & Probability Letters*, vol. 4, pp. 53-56, 1986.

[30] S. Hettich, C.L. Blake and C.J. Merz: UCI Repository of machine learning databases; http://www.ics.uci.edu/~mlearn/MLRepository.html University of California, Irvine, Dept. of Information and Computer Sciences, 1998.

[31] S. Miyamoto and S. Umayahara, "Fuzzy clustering by quadratic regularization," *Proc. 1998 IEEE Int. Conf. Fuzzy Syst.*, pp.1394-1399, 1998.

[32] M. Yasuda, T. Furuhashi, M. Matsuzaki and S. Okuma, "Fuzzy clustering using deterministic annealing method and its statistical mechanical characteristics," *Proc. 2001 IEEE Int. Conf. Fuzzy Syst.*, vol.3, pp.797-800, 2001.



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.

He is currently a Research Associate, Department of Computer Sciences and Intelligent Systems, Osaka Prefecture University. His research interests include hybrid techniques of fuzzy clustering and multivariate analysis, data mining with fuzzy data analysis and neural networks. He received paper award and young investigator award from Japan Society for Fuzzy Theory and Intelligent Informatics (SOFT) in 2002 and 2005, respectively, and gave a tutorial on "Introduction to Clustering Techniques" at 2004 IEEE Int. Conf. Fuzzy Systems (FUZZ-IEEE 2004).

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 of Matsushita Electric Industrial Co., Ltd., Tokyo, Japan. From 1981 to 1993, he was a Research Associate, Assistant Professor, and Associate Professor at Osaka Prefecture University, where he is currently a Professor in the Department of Computer Sciences and Intelligent Systems. His fields of interest are adaptive modeling of GMDH-type neural networks, fuzzy c-means clustering and classifier, data mining with fuzzy data analysis, human-machine interface, and cognitive engineering.

Francesco Masulli (M'91-SM'04) is an Associate Professor of Computer Science at the University of Genova (Italy). He held previous appointments at the Italian National Institute for Nuclear Physics, Ansaldo Automazione Co., the University of Genova and the University of Pisa, and has been also a visiting scientist at University of Nijmegen (Holland) and at the International Computer Science Institute in Berkeley (California).

He published more than 120 scientific papers in Machine Learning, Neural Networks, Fuzzy Systems and Pattern Recognition. He co-edited the books "Neural Networks in Biomedicine" (1994), "New Trends in Fuzzy Logic" (1996), "Advances in Fuzzy Systems and Intelligent Technologies" (2000), "Soft Computing Applications" (2002), and "Fuzzy Logic and Applications" (2006), and serves as a co-chair of the SIG Bioinformatics of the International Neural Network Society (INNS) and as an Associate Editor the international journal "Intelligent Automation and Soft Computing".

Stefano Rovetta received a Laurea degree in Electronic Engineering and a PhD degree in Models, Methods and Tools for Electronic and Electromagnetic Systems, both from the University of Genova.

He is currently assistant professor in computer science. Since 1992 he has been doing research on neural network models, implementations and applications in biomedical data processing, image and video analysis and processing, and multimedia information processing. His current research interests include machine learning and clustering techniques for high-dimensional biomedical data analysis and document analysis. He is the responsible for the Genova University in a technologic and cultural exchange project between the European Union and India, funded by the European Commission, and the responsible for a research activity on clustering in medical image analysis, funded by the Italian Ministry of Education, University and Research. He also took part to many funded research projects and produced more than 80 international scientific publications. Associate researcher of the former National Institute for the Physics of Matter, member of the National Institute of Higher mathematics "Francesco Severi", member of the international working group on "Soft Computing in Image Processing".