Robust Local Principal Component Analyzer with Fuzzy Clustering *

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

Abstract

Non-linear extensions of Principal Component Analysis (PCA) have been developed for detecting the lower-dimensional representations of real world data sets and local linear approaches are used widely because of their computational simplicity and understandability. Fuzzy c-Varieties (FCV) is the linear fuzzy clustering algorithm that estimates local principal component vectors as the vectors spanning prototypes of clusters. Least squares techniques, however, often fail to account for "outliers", which are common in real applications. In this paper, we propose a technique for making the FCV algorithm robust to intra-sample outliers. The objective function based on the lower rank approximation of the data matrix is minimized by a robust M-estimation algorithm that is similar to FCM-type iterative procedures.

1 Introduction

Principal Component Analysis (PCA) is a popular technique for learning low-dimensional linear models from multivariate data sets and its various applications include feature extraction, image processing, dimension reduction, object modeling and so on. However, the linear models based on least squares techniques are sensitive to outliers and the derived models are easily influenced by noise. In real applications, we often suffer from two different types of noise. One is the case that the data set includes noise samples and we must eliminate the whole elements of the noise samples. The other is the data including intra-sample outliers. For example, in the case of face recognition, all sample data must be face images and other (noise) pictures should be removed. On the other hand, if only a part of a image is blotted, we should analyze the data set not by removing the image but by ignoring only the noise pixels. In this way, intra-sample noise is also common in real world applications. For the robust principal component analysis of a noisy data set including intra-sample outliers, De la Torre *et al.* [1, 2] proposed a robust subspace learning technique based on robust M-estimation and applied the technique to a problem in computer vision by modeling outliers that typically occur at the pixel level.

In the knowledge discovery from real world databases, however, single linear models are often too simple to reveal the detailed features of the high-dimensional data sets. Then, many researchers are trying to extend the linear models into mixed models that can approximate the local linear features of non-linearly distributed data sets [3, 4, 5]. The local modeling is a kind of non-linear data analysis and seems to be more useful than other non-linear techniques such as principal curves [6] and neural networks [7, 8] in points of computational simplicity and understandability. Fuzzy c-Varieties (FCV) [9, 10] is a linear fuzzy clustering technique that captures the local linear structures of data sets. The prototypes of clusters are lower dimensional linear varieties spanned by vectors forming orthnormal bases of subspaces. Because the basis vectors are derived by solving the eigenvalue problems of fuzzy scatter matrices, they are regarded as the local principal component vectors. In this sense, the FCV algorithm is a kind of local PCA technique.

The result of fuzzy clustering algorithm is also influenced by outliers and several techniques for handling noise have been proposed. Dave's Noise Clustering [11] introduced an additional "noise cluster" so that all the noise samples could be dumped into that single cluster and other clusters could capture the local structures ignoring the noise samples. In the noise clustering approach, the sum of the memberships

*IJCNN 2003 Conference Proceedings, 732-737 (2003)

of a sample for the good clusters is constrained to be smaller than 1. The possibilistic clustering technique proposed by Krishnapuram *et al.* [12] tried to make the data partitioning robust by using a possibilistic constraint. In the possibilistic approach, there are no constraints on the memberships other than the requirement that they should be in [0,1], i.e., the membership value of a sample represents the possibility of the sample belonging to the cluster. However, in the analysis of large scale databases with high-dimensional observations, it is often the case that almost every sample includes a few noise elements and the conventional robust clustering methods fail to derive good results because they eliminate all the noise samples even though only a few elements of the samples are noise.

This paper proposes the technique for making the linear clustering algorithm robust by handling intra-sample noise. The FCV clustering is the same technique as the extraction of local principal components based on the minimization of the least squares criterion, which performs the lower rank approximation of the data matrix [13]. While the objective function of the FCV algorithm is composed of the distances between data points and prototypical linear varieties, we can derive the same function from the least squares criterion that achieves the component-wise approximation [14, 15]. Introducing the M-estimation technique, the lower rank approximation is performed by ignoring noise elements. The novel clustering algorithm is based on Iterative Reweighted Least Squares (IRLS) method [16] in which additional weight parameters enable to obtain the robust approximation by using iterative procedure without solving non-linear equation systems.

In numerical examples, we show the characteristic properties of our method.

2 Simultaneous Application of Robust Principal Component Analysis and Fuzzy Clustering

2.1 Local Principal Component Analysis Using Least Squares Criterion

Let $X = (x_{ij})$ denote an $(n \times m)$ data matrix consisting of m dimensional observation of n samples. In this paper, we often represent the data matrix as $X = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m)$ using n dimensional column vectors $\boldsymbol{x}_i$'s composed of the elements of the i-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. (In the following, bold symbols represent column vectors and the column vectors superscripted by "$\sim$" are composed of the row elements of a matrix.)

The goal of the simultaneous approach to PCA and fuzzy clustering is to partition the data set using local principal component vectors to express local linear structures. FCV is the clustering method that partitions a data set into C linear fuzzy clusters. The objective function of FCV consists of distances from data points to p dimensional prototypical linear varieties spanned by linearly independent vectors $\boldsymbol{a}_{cj}$'s as follows [9, 10]:

$$\begin{aligned} L_{fcv} &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}^{\theta}\Big\{(\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c)^\top(\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c) \\ &\quad -\sum_{j=1}^{p} \boldsymbol{a}_{cj}^\top R_{ci}\boldsymbol{a}_{cj}\Big\}, \end{aligned} \tag{1}$$

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

where u_{ci} denotes the membership degree of the data point $\boldsymbol{x}_i$ to the c-th cluster and $\top$ represent the transpose of the vector. $\boldsymbol{b}_c$ is the center of the c-th cluster. The weighting exponent θ is added for fuzzification. The larger θ is, the fuzzier the membership assignments are.

From the necessary condition for the optimality $\partial L_{fcv}/\partial \boldsymbol{a}_{cj} = \boldsymbol{0}$, the optimal $\boldsymbol{a}_{cj}$'s are derived by solving the following eigenvalue problem.

$$S_{fc}\boldsymbol{a}_{cj} = \mu_{cj}\boldsymbol{a}_{cj}, \tag{3}$$

where S_{fc} is the generalized fuzzy scatter matrix,

$$S_{fc} = \sum_{i=1}^{n} u_{ci}^{\theta} R_{ci}. \tag{4}$$

Because the optimal $\boldsymbol{a}_{cj}$'s are the eigenvectors corresponding to the largest eigenvalues, the vectors are regarded as the fuzzy principal component vectors extracted in each cluster considering the memberships [17].

In the same way, cluster centers and memberships are updated from the conditions $\partial L_{fcv}/\partial \boldsymbol{b}_c = \mathbf{0}$ and $\partial L_{fcv}/\partial u_{ci} = 0$ respectively. Iterative algorithm is used to derive the clustering result.

Honda *et al.* [14, 15] proposed to modify the objective function using least squares criterion and extended the FCV clustering algorithm to incomplete data sets including missing values. Introducing membership u_{ci}'s, the least squares criterion for local PCA is defined as

$$L_{lsc} = \sum_{c=1}^{C} \mathrm{tr}\Big\{(X - Y_c)^\top U_c^\theta (X - Y_c)\Big\}, \tag{5}$$

where $U_c = \mathrm{diag}(u_{c1}, \cdots, u_{cn})$. $Y_c = (y_{cij})$ denotes the lower rank approximation of the data matrix X in the c-th cluster,

$$Y_c = F_c A_c^\top + \mathbf{1}_n \boldsymbol{b}_c^\top, \tag{6}$$

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.

With fixed memberships, the extraction of local principal components in each cluster is equivalent to the calculation of F_c, A_c and $\boldsymbol{b}_c$ such that the least squares criterion of Eq.(5) is minimized.

From the necessary condition for the optimality of the objective function, $\partial L_{lsc}/\partial \boldsymbol{b}_c = \mathbf{0}$, we have

$$\boldsymbol{b}_c = (\mathbf{1}_n^\top U_c^\theta \mathbf{1}_n)^{-1}(X^\top - A_c^\top F_c^\top) U_c^\theta \mathbf{1}_n, \tag{7}$$

and if $F_c^\top U_c^\theta \mathbf{1}_n = \mathbf{0}$,

$$\boldsymbol{b}_c = (\mathbf{1}_n^\top U_c^\theta \mathbf{1}_n)^{-1} X^\top U_c^\theta \mathbf{1}_n. \tag{8}$$

Here, Eq.(8) is equivalent to the updating rule for the cluster center $\boldsymbol{b}_c$ in the FCV algorithm. Substituting Eqs.(6) and (8), Eq.(5) is

$$\begin{aligned} L_{lsc} = & \sum_{c=1}^{C} \Big\{ \mathrm{tr}(X_c^\top U_c^\theta X_c) - 2\mathrm{tr}(X_c^\top U_c^\theta F_c A_c^\top) \\ & + \mathrm{tr}(A_c F_c^\top U_c^\theta F_c A_c^\top) \Big\}, \end{aligned} \tag{9}$$

where $X_c = X - \mathbf{1}_n \boldsymbol{b}_c^\top$.

From $\partial L_{lsc}/\partial F_c = O$,

$$F_c A_c^\top A_c = X_c A_c. \tag{10}$$

Under the condition that $A_c^\top A_c = I$, we have $F_c = X_c A_c$ and the objective function is transformed as follows:

$$\begin{aligned} L_{lsc} &= \sum_{c=1}^{C} \Big\{ \mathrm{tr}(X_c^\top U_c^\theta X_c) - \mathrm{tr}(A_c^\top X_c^\top U_c^\theta X_c A_c) \Big\} \\ &= L_{fcv}. \end{aligned} \tag{11}$$

Therefore it can be said that Eq.(5) is equivalent to the objective function of FCV and the minimization problem is solved by computing the p largest singular values of the fuzzy scatter matrix and their associated vectors.

2.2 Robust Local Principal Component Analysis

When we deal with a data matrix including noise elements, the local models based on least squares techniques are easily distorted. In this section, we introduce the technique for handling intra-sample outliers to local PCA. The technique of M-estimation is the useful method for estimating robust models. The goal of the method is to derive the solution ignoring outliers that don't conform to the assumed statistical model.

To estimate the robust subspace from a data set including intra-sample noise, De la Torre *et al.* [1, 2] proposed to minimize the following robust energy function:

$$L_{rpca} = \sum_{i=1}^{n}\sum_{j=1}^{m}\rho(x_{ij} - \sum_{k=1}^{p} f_{ik}a_{jk} - b_j), \tag{12}$$

where $\rho(\cdot)$ is a class of robust ρ-functions [16]. In [1, 2], they used the Geman-McClure error function [18],

$$\rho(x) = \frac{x^2}{x^2 + \sigma_j^2}, \tag{13}$$

where σ_j is a scale parameter that controls the convexity of the robust function. In the process of the optimization, the value of σ_j is decreased by the deterministic annealing technique. To solve the minimization problem, they proposed to use not only the iteratively reweighted least-squares (IRLS) technique [16] but also the gradient descent method with a local quadratic approximation. In the following, we propose a robust local PCA technique by introducing the ρ-function into the FCV algorithm with the least squares criterion. The objective function of Robust FCV with entropy regularization [19] is defined as follows:

$$\begin{aligned} L_{rfcv} &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\sum_{j=1}^{m}\rho(x_{ij} - \sum_{k=1}^{p} f_{cik}a_{cjk} - b_{cj}) \\ &\quad + \lambda\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci}. \end{aligned} \tag{14}$$

The entropy term is added for fuzzification instead of the weighting exponent in the standard FCV algorithm. The larger λ is, the fuzzier the membership assignments are. The fuzzification technique has several merits, e.g., "singularities" don't occur even if several sample points are on the prototypes and cluster centers are the means of $\tilde{\boldsymbol{x}}_i$ simply weighted by u_{ci}'s.

To obtain a unique solution, the objective function is minimized under the constraints that

$$F_c^\top U_c F_c = I \quad ; \quad c = 1, \cdots, C, \tag{15}$$

$$F_c^\top U_c \mathbf{1}_n = \mathbf{0} \quad ; \quad c = 1, \cdots, C, \tag{16}$$

$$\sum_{c=1}^{C} u_{ci} = 1 \quad ; \quad i = 1, \cdots, n, \tag{17}$$

and $A_c^\top A_c$ is orthogonal.

Here, the optimal solution cannot be derived from eigenvalue problems because the clustering criterion is transformed by the non-linear ρ function. In this paper, we derive the solution based on the IRLS technique in which the minimization problem is formulated as a weighted least squares problem with an $(n \times m)$ weight matrix $W_c = (w_{cij})$ in each cluster. w_{cij} represents the positive weight for the previous residual,

$$e_{cij} = x_{ij} - \sum_{k=1}^{p} f_{cik}a_{cjk} - b_{cj}. \tag{18}$$

For the Geman-McClure ρ function,

$$w_{cij} = \frac{\psi(e_{cij}, \sigma_j)}{e_{cij}}, \tag{19}$$

where

$$\begin{aligned} \psi(e_{cij}, \sigma_j) &= \frac{\partial\rho(e_{cij})}{\partial e_{cij}} \\ &= \frac{2e_{cij}\sigma_j^2}{(e_{cij}^2 + \sigma_j^2)^2}. \end{aligned} \tag{20}$$

To derive the optimal A_c and $\boldsymbol{b}_c$, we rewrite Eq.(14) as follows:

$$
\begin{aligned}
L_{rfcv} &= \sum_{c=1}^{C}\sum_{j=1}^{m}(\boldsymbol{x}_j - F_c\tilde{\boldsymbol{a}}_{cj} - \mathbf{1}_n b_{cj})^\top U_c W_{cj} \\
&\quad \times(\boldsymbol{x}_j - F_c\tilde{\boldsymbol{a}}_{cj} - \mathbf{1}_n b_{cj}) \\
&\quad +\lambda\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci}, \qquad (21)
\end{aligned}
$$

where

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

$$W_{cj} = \mathrm{diag}(w_{c1j}, \cdots, w_{cnj}).$$

From $\partial L_{rfcv}/\partial\tilde{\boldsymbol{a}}_{cj} = \mathbf{0}$ and $\partial L_{rfcv}/\partial b_{cj} = 0$, we have

$$\tilde{\boldsymbol{a}}_{cj} = (F_c^\top U_c W_{cj} F_c)^{-1} F_c^\top U_c W_{cj}(\boldsymbol{x}_j - \mathbf{1}_n b_{cj}), \qquad (22)$$

$$b_{cj} = (\mathbf{1}_n^\top U_c W_{cj}\mathbf{1}_n)^{-1}\mathbf{1}_n^\top U_c W_{cj}(\boldsymbol{x}_j - F_c\tilde{\boldsymbol{a}}_{cj}). \qquad (23)$$

In the same way, we can derive the optimal F_c and u_{ci}. Eq.(14) is equivalent to

$$
\begin{aligned}
L_{rfcv} &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}(\tilde{\boldsymbol{x}}_i - A_c\tilde{\boldsymbol{f}}_{ci} - \boldsymbol{b}_c)^\top \tilde{W}_{ci} \\
&\quad \times(\tilde{\boldsymbol{x}}_i - A_c\tilde{\boldsymbol{f}}_{ci} - \boldsymbol{b}_c) \\
&\quad +\lambda\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci}, \qquad (24)
\end{aligned}
$$

and $\partial L_{rfcv}/\partial\tilde{\boldsymbol{f}}_{ci} = \mathbf{0}$ and $\partial L_{rfcv}/\partial u_{ci} = 0$ yields

$$\tilde{\boldsymbol{f}}_{ci} = (A_c^\top \tilde{W}_{ci} A_c)^{-1} A_c^\top \tilde{W}_{ci}(\tilde{\boldsymbol{x}}_i - \boldsymbol{b}_c), \qquad (25)$$

$$
\begin{aligned}
u_{ci} &= \exp\Big\{-(\tilde{\boldsymbol{x}}_i - A_c\tilde{\boldsymbol{f}}_{ci} - \boldsymbol{b}_c)^\top \tilde{W}_{ci} \\
&\qquad \times(\tilde{\boldsymbol{x}}_i - A_c\tilde{\boldsymbol{f}}_{ci} - \boldsymbol{b}_c)/\lambda - 1\Big\}, \qquad (26)
\end{aligned}
$$

where

$$\tilde{W}_{ci} = \mathrm{diag}(w_{ci1}, \cdots, w_{cim}). \qquad (27)$$

The proposed algorithm can be written as follows.

Robust Fuzzy c-Varieties (Robust FCV) Algorithm

Step 1 Initialize $U_c, A_c, \boldsymbol{b}_c, F_c$ randomly in each cluster and normalize them so that they satisfy the constraints Eqs.(15)-(17) and $A_c^\top A_c$ is orthogonal.

Step 2 Calculate the initial W_c in each cluster.

Step 3 Update A_c's using Eq.(22) and transform them so that each $A_c^\top A_c$ is orthogonal.

Step 4 Update F_c's using Eq.(25) and normalize them so that they satisfy the constraints Eqs.(15) and (16).

Step 5 Update $\boldsymbol{b}_c$'s using Eq.(23).

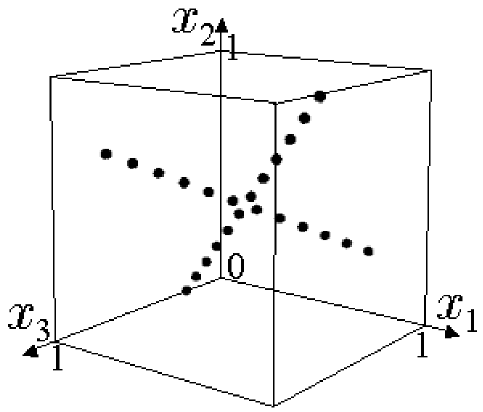


Figure 1: 3-D Plots of Original Data Set

Step 6 Update U_c's using Eq.(26) and normalize them so that Eq.(17) holds.

Step 7 If

$$\max_{c,i} \mid u_{ci}^{NEW} - u_{ci}^{OLD} \mid < \epsilon_1,$$

then go to Step 8. Otherwise, return to Step 3.

Step 8 Update W_c's. If

$$\max_{c,i,j} \mid w_{cij}^{NEW} - w_{cij}^{OLD} \mid < \epsilon_2,$$

then stop. Otherwise, return to Step 3.

2.3 Handling Missing Values

Not only noise or outliers but also missing values are common in real world data analysis. In such cases, a priori knowledge about the observation is often available while many noise elements aren't identified before the analysis. In [14, 15], missing values in data matrix are ignored by multiplying "0" weights to the corresponding reconstruction errors.

The novel method is also applied to incomplete data sets. To handle missing values, the weight parameter w_{cij} is redefined as follows:

$$w_{cij} = \begin{cases} \frac{2\sigma_j^2}{(e_{cij}^2+\sigma_j^2)^2} & ; x_{ij} \text{ is observed.} \\ 0 & ; x_{ij} \text{ is missing.} \end{cases} \tag{28}$$

If all the weight w_{cij} for the observed element x_{ij} is 1, the proposed method is equivalent to the FCV algorithm with missing values [14].

3 Numerical Experiments

3.1 Analysis of Artificial Data Sets

We performed numerical experiments using an artificial data set shown in Fig. 1. The 3-D data set composed of 24 samples forms two lines and the goal of the analysis is to capture these lines. Applying the FCV algorithm ($\theta = 2$) to this complete data set, we could partition the samples into two linear clusters whose prototypes are shown in Table 1. The prototypes represented the two lines properly.

First, we compared the robustness of our novel method with the conventional noise clustering technique. Replacing randomly selected elements with noise values, we made an noisy data set that includes 21% noise elements. Fig. 2 and Fig. 3 show the 3-D plots and the 2-D projections of the noisy data set. Note that each "noise sample" includes only one noise element, i.e., the noise sample points are on

Table 1: Clustering Result by FCV with Complete Data

		$\boldsymbol{b}_c$			$\boldsymbol{a}_c$		
		x_1	x_2	x_3	x_1	x_2	x_3
FCV	$c = 1$	0.50	0.50	0.50	-0.41	0.41	0.82
	$c = 2$	0.50	0.50	0.50	0.67	0.67	0.33

Table 2: Clustering Result with Noisy Data

		$\boldsymbol{b}_c$			$\boldsymbol{a}_c$		
		x_1	x_2	x_3	x_1	x_2	x_3
FCV	$c = 1$	0.52	0.48	0.50	-0.43	0.41	0.81
	$c = 2$	0.49	0.46	0.50	0.71	0.64	0.31
FCV	$c = 1$	0.49	0.49	0.50	-0.42	0.43	0.80
(NC)	$c = 2$	0.49	0.47	0.49	0.67	0.67	0.33
RFCV	$c = 1$	0.47	0.52	0.55	-0.41	0.42	0.81
	$c = 2$	0.50	0.48	0.50	0.67	0.67	0.33

Table 3: Clustering Result with Incomplete Noisy Data

		$\boldsymbol{b}_c$			$\boldsymbol{a}_c$		
		x_1	x_2	x_3	x_1	x_2	x_3
FCV	$c = 1$	0.54	0.46	0.52	-0.47	0.53	0.71
	$c = 2$	0.42	0.46	0.56	0.67	0.65	0.35
FCV	$c = 1$	0.44	0.43	0.54	-0.43	0.56	0.71
(NC)	$c = 2$	0.50	0.46	0.49	0.68	0.66	0.33
RFCV	$c = 1$	0.50	0.49	0.49	-0.42	0.42	0.80
	$c = 2$	0.53	0.51	0.51	0.67	0.66	0.33

the line in one of three projections in Fig. 3. Table 2 shows the prototypes derived by the standard FCV algorithm, the FCV algorithm with noise clustering mechanism and our proposed method. For the comparison, the principal component vectors of our method are normalized to be unit length. $\lambda = 0.05$ and the scale parameter σ_j was annealed by the following schedule:

$$\sigma_j^2 = \frac{0.5}{\log(t+2)}, \tag{29}$$

where t is the iteration index. Fig.4 shows the trajectory of the coefficient. Although the prototypes of the standard FCV algorithm were influenced by outliers, the noise clustering version and the novel algorithm could capture the two lines properly. Here, the noise clustering ignored the samples including noise element and assigned them into noise cluster. On the other hand, our novel algorithm ignored only the noise elements and assigned the noise samples into proper clusters. In this way, our algorithm provides a similar result with noise clustering when the data set includes only a few noise elements.

Second, we performed the same experiment with an incomplete data set including not only noise elements but also missing values. The data set was made by withholding one element from 10 samples. $\circ$'s in Fig. 3 represent the samples including a missing value. For example, $\circ$ in the projection on $x_1 - x_2$ field indicates that the third element of the sample is missing. In the conventional methods, the analysis of incomplete data is impossible without eliminating all samples with missing values. Then, the standard FCV algorithm and the FCV with noise clustering algorithm were performed after the preprocessing. Table 3 shows the comparison of the prototypes of the derived clusters. Only our proposed method could analyze the data set without both the influences of noise and the loss of the information.

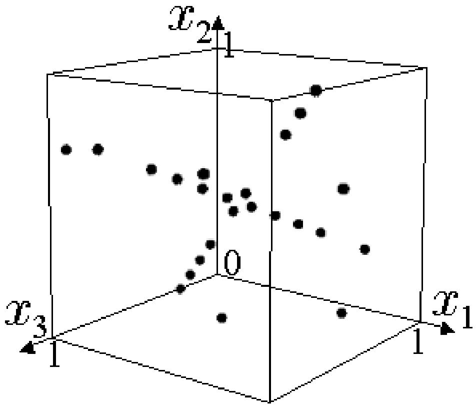
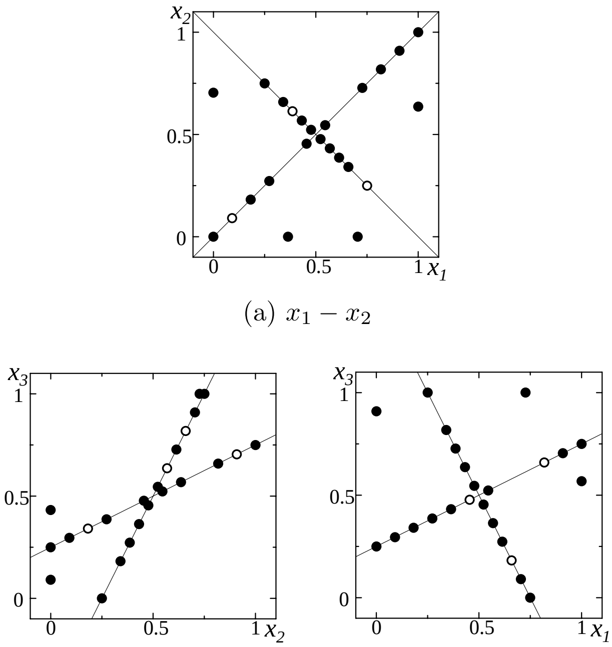


Figure 2: 3-D Plots of Noisy Data Set



(a) $x_1 - x_2$

(b) $x_2 - x_3$

(c) $x_1 - x_3$

Figure 3: 2-D Projections of Noisy Data Set

3.2 Application to Collaborative Filtering

Local linear models extracted from incomplete data sets can be also applied to the missing value estimation because the approximation matrix Y_c includes no missing elements. Collaborative filtering is an attractive application of the missing value estimation and linear fuzzy clustering technique was applied to the task [15]. We implemented the novel technique presented in the previous section for the collaborative filtering and tested them with the ratings data collected for purposes of anonymous review from the MovieLens movie recommendation site [20]. The data set is composed of 100,000 ratings from 943 users and every user evaluated at least 20 ratings on a scale from 1 to 5 based on the semantic differential (SD) method. 20,000 ratings were randomly selected to be the test data. Here, we used only 1,240 movies that were evaluated by at least 4 users because other movies' ratings were difficult to predict considering the correlations. In proposed technique, we partitioned the users into two clusters and extracted one principal component vector in each cluster. The parameters were given by $\lambda = 6.0$ and

$$\sigma_j^2 = \frac{5.0}{\log(t+2)}. \tag{30}$$

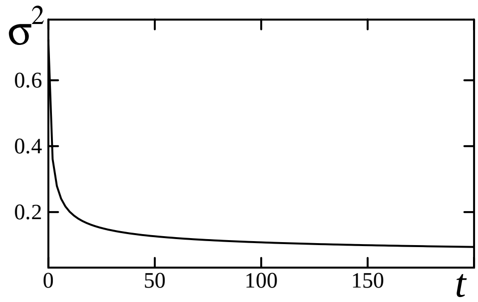


Figure 4: Annealing Schedule

Table 4: Comparison of Prediction Algorithms

Algorithm	**MAE**	**ROC**
Non-personalized Method	0.821	0.714
GroupLens	0.762	0.762
FCV with Missing Values	0.754	0.777
Robust FCV	0.751	0.789

The missing values were predicted by using corresponding elements of the approximation matrix Y_c in the cluster where the user had the maximum membership. It means that we estimated the missing values based on the assumption that the data points including missing values should exist on the nearest point to linear varieties spanned by local principal component vectors. In addition, we also predicted the ratings using original GroupLens [21], a non-personalized prediction method [22] and the FCV with missing values [15]. In the non-personalized prediction method, we computed the ratings using deviation-from-mean average over all users. Table 4 compares the result of them.

For assessing the accuracy of the three prediction methods, we used not only the mean absolute error (MAE), but also the receiver operating characteristic (ROC) sensitivity. ROC sensitivity is a measure of the diagnostic power of a system [23]. The sensitivity refers to the probability of a randomly selected good item being accepted by the filter. The greater the value is, the richer the performance becomes. The maximum value is one. In this paper, the movies whose ratings were larger than 3 were regarded as good items and the filtering system recommended the movies whose predictions were larger than 3.5. The proposed method provided the best performance, i.e. the local linear models derived by the proposed method represented the principal local features properly. In this way, the proposed method is useful to extract robust local features of large scale databases.

4 Conclusion

In this paper, we proposed a new linear fuzzy clustering method that is robust to intra-sample noise. Introducing the "component-wise" robust M-estimation technique, the FCV algorithm is generalized to noisy data sets. Although the objective function includes non-linear functions, additional weight parameters make it possible to derive the solution by an FCM-like iterative algorithm in which the weight parameters control the responsibility of each element for the local modeling. While the conventional noise clustering techniques or the possibilistic approaches ignore all "noise samples" even when the samples include only a few noise elements, the proposed method ignores only "noise elements". When the a priori information about the observation is available, we can also handle missing values by constraining the associated weights to be 0.

References

[1] F. de la Torre, and M. J. Black, "Robust principal component analysis for computer vision," *Proc. of International Conference on Computer Vision*, pp. 362-369, 2001.

[2] F. de la Torre, and M. J. Black, "A framework for robust subspace learning," *International Journal of Computer Vision*, to appear.

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

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

[6] T. Hastie, and W. Stuetzle, "Principal curves," *J. of American Statistical Association*, vol. 84, pp. 502-516, 1989

[7] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE Journal*, vol. 37, no. 2, pp. 233–243, 1991

[8] E. Oja, "Data compression, feature extraction, and autoassociation in feedforward neural networks," *Artificial Neural Networks*, Elsevier Science Publishers, pp. 737-745, 1991

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

[10] 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.

[11] R. N. Dave, "Characterization and detection of noise in clustering," *Pattern Recognition Letters*, vol. 12, no. 11, pp. 657-664, 1991.

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

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

[14] K. Honda, N. Sugiura, and H. Ichihashi, "Simultaneous approach to principal component analysis and fuzzy clustering with missing values," *Proc. of Jnt. 9th Int. Fuzzy Syst. Assoc. World Cong. and 20th North American Fuzzy Inf. Processing Soc. Int. Conf.*, pp. 1810-1815, 2001

[15] 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

[16] 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.

[17] Y. Yabuuchi, and J. Watada, "Fuzzy principal component analysis and its application," *Biomedical Fuzzy and Human Sciences*, vol. 3, no. 1, pp. 83-92, 1997.

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

[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] MovieLens Web Page; http://www.movielens.org/

[21] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "Grouplens: an open architecture for collaborative filtering of netnews," *Proc. of ACM Conference on Computer-Supported Cooperative Work*, pp. 175-186, 1994.

[22] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, "An algorithmic framework for performing collaborative filtering", *Proc. of Conference on Research and Development in Information Retrieval*, 1999.

[23] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, no. 4857, pp. 1285-1289, 1988.