# Multi-Manifold Clustering

Yong Wang[1,2], Yuan Jiang[2], Yi Wu[1], and Zhi-Hua Zhou[2]

[1] Department of Mathematics and Systems Science
National University of Defense Technology, Changsha 410073, China
yongwang82@gmail.com, wuyi_work@sina.com
[2] National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
{jiangy, zhouzh}@lamda.nju.edu.cn

**Abstract.** Manifold clustering, which regards clusters as groups of points around compact manifolds, has been realized as a promising generalization of traditional clustering. A number of linear or nonlinear manifold clustering approaches have been developed recently. Although they have attained better performances than traditional clustering methods in many scenarios, most of these approaches suffer from two weaknesses. First, when the data are drawn from hybrid modeling, i.e., some data manifolds are separated but some are intersected, existing approaches could not work well although hybrid modeling often appears in real data. Second, many approaches require to know the number of clusters and the intrinsic dimensions of the manifolds in advance, while it is hard for the user to provide such information in practice. In this paper, we propose a new manifold clustering approach, mumCluster, to address these issues. Experimental results show that the performance of the proposed mumCluster approach is encouraging.

## 1 Introduction

Traditional clustering methods, such as $K$-means [1], are based on the idea that a cluster is centered around a single point when measuring similarity. Recently, a large number of research efforts have shown that the perceptually meaningful structure of the points possibly resides on a low-dimensional manifold [2, 3]. Therefore, regarding cluster as a group of points around a compact manifold becomes a reasonable and promising generalization of traditional clustering, leading to *manifold clustering* [4].

Roughly speaking, the research on manifold clustering can be classified into two branches, i.e., linear and nonlinear. Generalized Principal Component Analysis (GPCA) [5, 6] and $K$-planes [7–9] assume the samples to be well approximated by a mixture of affine subspaces (or linear manifolds). However, manifolds in natural data are generally nonlinear in the original space [2]. Spectral clustering (SC) [10, 11] is a good option when the samples are lying on separated clusters where each cluster contains points sampled from a single nonlinear manifold. Alternatively, Cao and Haralick [12] use the local dimension and mean square error to infer clusters. However, when there are intersections among clusters, their performance will degenerate. $K$-manifolds [4] is primarily motivated to cluster samples generated from intersecting nonlinear manifolds, which will fail when the clusters are widely separated.

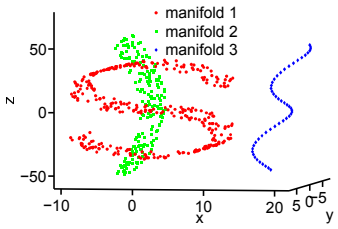



**Fig. 1.** Data points drawn from a hybrid modeling.

There are two main difficulties for existing methods. On the one hand, they usually work well either in separated case or in intersecting case. When the input data points are drawn from a hybrid modeling (see Figure 1) where some manifolds are separated, while some others are intersected with each other, the quality of clustering degenerate. On the other hand, many of existing methods require the user to provide the number of clusters and their intrinsic dimensions in advance, while such information are difficult to be given in practice. For example, considering a data set consisting of face images of different individuals under various lighting conditions, it is difficult for the user to know whether the underlying manifolds are separated or intersected, as well as the number of clusters and the intrinsic dimensions ahead. Thus, to enable manifold clustering to deal with more real tasks, it is important to design manifold clustering approaches which are able to work well when the samples are drawn from hybrid modeling, and which can adaptively determine the number of clusters and dimensions.

In this paper, we propose a new manifold clustering method called mumCluster (MUlti-Manifold Clustering). Our basic idea is based on the observation that if we can make the constructed undirected graph in spectral clustering more faithful, i.e., data points belonging to different manifolds will not be connected, then spectral clustering can be used to identify different manifolds accurately. Thus, our scheme first identifies the separate subsets of the original data, and then determines whether a subset is composed of a single manifold or intersecting manifolds. For each intersecting subset, we will exclude the influence of the inaccurate connected relationships among different manifolds. Finally, spectral clustering is used to further infer clusters. Moreover, a strategy is developed to automatically determine the number of manifold clusters and their corresponding dimensions.

The rest of this paper is organized as follows: Section 2 briefly reviews the related manifold clustering methods. In Section 3, the mumCluster method is presented, followed by a strategy to determine the number of clusters and their dimensions. Computational complexity analysis of the proposed method is also presented in this section. In Section 4, we experimentally evaluate the performance of our proposed method using synthetic and real-world data. Section 5 concludes this paper.

## 2 Related Work

Cluster analysis [13] seeks to group internally similar objects into the same cluster while dissimilar objects into different clusters. Traditional clustering methods, such as

$K$-means [1], assume the data are centered around some prototypes. They could not separate clusters that are nonlinearly separable or centered around manifolds.

GPCA [5, 6] and $K$-planes [7–9] are representative linear manifold clustering methods. GPCA models the underlying manifolds with a set of homogeneous polynomials, then the constructed models are used to infer clusters. Alternatively, $K$-planes addresses linear manifold clustering by iterating between assigning data to manifolds, and modeling a manifold to each cluster. Although successful for mixtures of linear clusters, both of them fail to deliver good performance in the presence of nonlinear structures (e.g., Figure 3 (a) and (b)). Since nonlinear methods can also work well on linear clusters, in this paper, we focus on the nonlinear manifold clustering.

Spectral clustering [10, 11] is a good option for nonlinear manifold clustering when samples are generated from separated clusters where each cluster contains data points from a single manifold [14]. However, when there are intersections in some areas, spectral clustering could not work well (e.g., Figure 3 (c)). The reason is that the performance of spectral clustering is heavily relied on the constructed undirected graph, different clusters near a manifold intersection will easily be connected by the undirected graph, thus diffusing information across the wrong manifolds [15]. $K$-manifolds [4] groups data lying on intersecting nonlinear manifolds, which begins by estimating geodesic distances between points, then an expectation maximization (EM) type strategy is used to iterate between estimating the manifolds using node-weighted MDS and assigning each point to the specified manifolds. Unfortunately, the estimation of geodesic distances fails when there are separated clusters, leading to incorrect clustering (e.g., Figure 3 (d)). The method most related to ours was proposed by Cao and Haralick [12], which groups neighboring points into a cluster if they have the same local dimension and the mean square error of representing the new cluster is small. This method can handle the hybrid modeling to some extent, by using graph methods to identify different connected components. However, it is primarily based on the local dimension, thus the method usually treats the intersections as clusters since the local dimension in the intersections are higher than the other areas (e.g., Figure 3 (e)).

## 3 MumCluster

Given a set of data points $X = \{x_i \in \Re^D, i = 1, 2, \cdots, N\}$ sampled from $k > 1$ distinct manifolds $\{\Omega_j \subseteq \Re^D, j = 1, 2, \cdots, k\}$ with dimension $d_j = \dim(\Omega_j)$, $0 < d_j < D$. The samples are unorganized, i.e., we do not know which points belong to which manifold. Moreover, some manifolds are intersected with each other which form intersecting manifolds. Our objectives are:
1. *Identify the number of manifolds* $k$ *and their intrinsic dimensions* $\{d_j, j = 1, 2, \cdots, k\}$;
2. *Partition the given samples into the manifold(s) they belong to.*

Though a considerable amount of work has been done in this field, as we have reviewed before, they could not work well on the hybrid modeling. Moreover, many of them need the user to specify $k$ and $\{d_j, j = 1, 2, \cdots, k\}$. In what following, we propose the mumCluster method to address these issues.

Our main strategy is trying to construct more faithful undirected graph in spectral clustering, i.e., data points belonging to different manifolds will not be connected.

Therefore, mumCluster designs a "divide and conquer" strategy to realize this purpose. This scheme first divides the complicated intersecting manifolds from the single manifolds, then each intersecting subset is further divided into intersection areas and non-intersection areas. More attention is paid to the intersection areas, where many of the inaccurate connected relationships situated. The details of the method are presented in Subsection 3.1, followed by a strategy to automatically determine the number of clusters and their dimensions in Subsection 3.2. Complexity analysis is presented in Subsection 3.3.

### 3.1 To Deal with Hybrid Modeling

Generally, hybrid modeling can be divided into different connected subsets, with some subsets containing only single manifold, while the others containing intersecting manifolds. To deal with the two different structures separately, we propose to use spectral clustering to partition the samples coarsely into different connected subsets. Generally, there are different versions of spectral clustering. Following von Luxburg's suggestion [14], the following unsymmetrical normalized spectral clustering [10] is adopted:

1. Constructing a similarity graph $G$: Put an edge between node $i$ and $j$ if $i$ is among $L$ nearest neighbors of $j$, and vice versa.

2. Determining the weighted matrix $W$: If node $i$ and $j$ are connected, then put a weight $w_{ij}$ as $w_{ij} = 1$ (simple weight); otherwise, put $w_{ij} = 0$.

3. Spectral decomposition: Compute the first $r$ eigenvectors $u_1, u_2, \cdots, u_r$, corresponding to the $r$ smallest eigenvalues, of the generalized eigenproblem $Eu = \lambda Fu$, where $F$ is a diagonal matrix with $F_{ii} = \sum_j w_{ij}$ and $E = F - W$. Let $U = [u_1, u_2, \cdots, u_r] \in \Re^{N \times r}$.

4. Clustering by $K$-means: Group the points $y_i, i = 1, 2, \cdots, N$ into $r$ clusters using $K$-means, where $y_i$ is the vector corresponding to the $i$-th row of $U$.

In the above procedure, $r$ should be provided. We will discuss on how to decide $r$ in the next subsection.

After the different connected subsets $\Xi_c, c = 1, \cdots, r$ have been identified, the problem is how to determine their structure, i.e., single or intersecting. For this purpose, our basic idea is to resort to the intrinsic dimension $id$. It is based on the observation that if samples come from a single manifold, then the intrinsic dimension of each point on this manifold should be the same; otherwise, they are different. Details on estimating $id$ will be presented in the next subsection.

If the connected subset consists of a single manifold, then a manifold cluster has been revealed. However, for each intersecting subset $\Xi^{is}$, further procedures are needed to reveal different manifold clusters. The first should be to identify the intersection areas $\Pi^{ia}$ and the non-intersection areas $\Pi^{nia}$. Generally, the points in $\Pi^{ia}$ have higher dimension than the other parts. Therefore, the points with the highest dimension $d_{\max}$ should be first grouped into $\Pi^{ia}$. In practice, the structure in the intersection area is usually complex. To ensure this area to be identified accurately, the $\varepsilon$-neighbors can be used. That is,

$$x \in \Pi^{ia}, \quad \text{if} \quad \left\|x - x^{ip}\right\|^2 < \varepsilon, \tag{1}$$

where $x^{ip}$ is any point with dimension $d_{\max}$. Finally, $\Xi^{is}$ is divided into $\Pi^{ia}$ and $\Pi^{nia}$.

The points in $\Pi^{ia}$ and $\Pi^{nia}$ may consist of many small clusters (called *intersection clusters* and *non-intersection clusters*, respectively), which should be grouped in order to tackle them separately. Generally, these clusters are unconnected, thus spectral clustering can still be used here to group them. If the dimensions on some non-intersection clusters are different, it implies that there may still exist some other intersection clusters with lower $d_{\max}$. Therefore, we should go back to identify these areas until there is no hidden intersection.

The intersection area implies that there are different manifolds passing across each other which should be revealed. Though, the manifold clusters are nonlinear, each intersection cluster can be considered as a mixture of manifolds with linear structure since it is a local area. Thus, $K$-planes can be adopted to reveal the different manifolds (named *fine clusters*) in each intersection cluster. Specifically, given the number of clusters $k^*$ and the dimensions $d_1^*, d_2^*, \cdots, d_{k^*}^*$.

1. Initialization: Assign each point to a cluster randomly to give an initial partition $\{C_1^*, C_2^*, \cdots, C_{k^*}^*\}$. Then, alternating between the following two steps until convergence.

2. Cluster update: Find a center $\mu_i^*$ and a set of bases $\Phi_i = [\varphi_1^i, \varphi_2^i, \cdots, \varphi_{d_i^*}^i]$ for cluster $C_i^*$ such that the reconstruction error is minimum.

3. Cluster assignment: For each point $x_m^*$ in the considered intersection cluster, determine the space $j$ such that

$$\begin{aligned}&(x_m^* - \mu_j^*)^T (I - \Phi_j \Phi_j^T)(x_m^* - \mu_j^*)\\ =& \min_{i=1,\cdots,k^*} (x_m^* - \mu_i^*)^T (I - \Phi_i \Phi_i^T)(x_m^* - \mu_i^*),\end{aligned} \tag{2}$$

where $I$ is an identity matrix. Then, $x_m^*$ is assigned to the $j$-th cluster $C_j^*$.

As indicated before, the constructed undirected graph for each intersecting subset may connect different manifolds, making the partition of samples into the manifold they belong to impractical. To reveal different manifolds, the connections between them should be cut out, and should be preserved among the same manifold. Since the unfaithful connections mainly come from the different fine clusters, we cut the connections among them, while connect all the points in the same fine cluster to preserve the manifold structure. Finally, a new undirected graph $G_{new}$ is obtained for each intersecting subset $\Xi^{is}$. Thus, spectral clustering is used to finally group points in each $\Xi^{is}$ into different manifold clusters.

### 3.2 To Determine the Number of Clusters and the Intrinsic Dimensions

Hereinbefore, we have shown our scheme to partition the given samples into the manifold they belong to. However, it is based on the given number of clusters and their intrinsic dimensions, and how to adaptively determine these parameters are not resolved. In the following, we propose to use eigengap, local intrinsic dimension estimator and a new bottom-up search procedure to address these issues.

First, as demonstrated in [14], the number of connected components $r$ in the adopted spectral clustering equals the multiplicity $r$ of the eigenvalue zero of the generalized

eigen-problem. Therefore, $r$ can be determined by using the *eigengap* heuristic. That is,

$$\text{if} \quad |\lambda_l - \lambda_{l-1}| \leq 10^{-6} < |\lambda_{l+1} - \lambda_l|, \quad \text{then} \quad r = l, \tag{3}$$

where $10^{-6}$ is used to replace zero to avoid numeric problem.

The intrinsic dimension $id$ of each point can be estimated by using a local dimension estimator. It is based on the observation that though the manifold structures are globally nonlinear, they are locally linear [3]. Moreover, it is known that the first $id$ largest eigenvalues of the covariance matrix are significantly higher than the others and thus can be used as an estimation to the intrinsic dimension, when the original data are sampled from an $id$-dimensional manifold [16]. In more detail, we can estimate the intrinsic dimension by:

1. Calculate the local covariance matrix: For each point $x_i$, find its $L$ nearest neighbors $x_i^1, \cdots, x_i^L$, then calculate the local covariance matrix

$$C_i = 1/L \sum\nolimits_{j=1}^{L} (x_i^j - \mu_i)(x_i^j - \mu_i)^T, \tag{4}$$

where $\mu_i = 1/L \sum_{j=1}^{L} x_i^j$ is the mean vector.

2. Intrinsic dimension estimation: Determine the sorted eigenvalues $\lambda_1^i \geq \cdots \geq \lambda_D^i$ of $C_i$.

$$\text{if} \quad \lambda_j^i/\lambda_1^i < 0.05 \leq \lambda_{j-1}^i/\lambda_1^i, \quad \text{then} \quad id = j - 1. \tag{5}$$

More challenging is to determine $k^*$ and $d_1^*, d_2^*, \cdots, d_{k^*}^*$ in the $K$-planes algorithm which is used to reveal fine clusters in each intersection cluster. Our solution is based on a bottom-up search strategy, which starts from the lowest dimension $d_{\min}$. Moreover, we can determine the possible dimensions and the number of clusters, which reduce the search space. First, let us introduce the following notion.

**Definition: Effective Dimension (ED) [17]**
*Given $k$ subspaces $\Omega = \bigcup_{i=1}^{k} \Omega_i$ in $\Re^D$ of dimension $d_i < D$, and $N_i$ sample points $X_i = \{x_i^j, j = 1, \cdots, N_i\}$ drawn from each subspace $\Omega_i$, the effective dimension is defined to be:*

$$ED(X, \Omega) \triangleq 1/N \sum\nolimits_{i=1}^{k} d_i(D - d_i) + 1/N \sum\nolimits_{i=1}^{k} N_i d_i. \tag{6}$$

Effective dimension $ED(X, \Omega)$ is the "average" numbers needed to assign to per sample of $X$. Generally, there could be many manifold structures $\Omega$ which can fit $X$, while the manifold structure that leads to the minimum ED normally corresponds to an "efficient" and hence "natural" interpretation of the data, see [17]. Formally, ED is low if the number of clusters and dimension of each cluster are small. Therefore, to faithfully fit the underlying manifold structure, we should search for the structure which minimizes ED among all possible structures under certain criterion. To be consist with the $K$-planes algorithm, the reconstruction error is a good choice.

To reduce the search space, the following observation is considered: the intersection clusters are crossed by different manifolds, moving continuously from the non-intersection clusters. Suppose an intersection cluster is connected with $m$ non-intersection

mumCluster($X$, $L$, $\varepsilon$, $\zeta_{\max}$)

**Input:**

$X$: $D \times N$ feature matrix
$L$: number of nearest neighbors
$\varepsilon$: threshold for determining the intersection area
$\zeta_{\max}$: maximum error threshold

**Process:**

```
Construct graph G with weighted matrix W
Group using spectral clustering on W with eigengap
for each connected subset
  Compute the intrinsic dimension id for each point
  if id's are the same
    Output this connected subset as a cluster
  else
    Construct a new graph G_new
    Group using spectral clustering on G_new
  endif
end
```

**Output:**

$\{C_1, C_2, \cdots, C_k\}$: the results of clustering

**Fig. 2.** Pseudo-code of the mumCluster method

clusters, then the dimensions of the non-intersection clusters imply the possible dimensions of the fine clusters, while the number of non-intersection clusters limits the number of fine clusters.

Our bottom-up strategy can be summarized as follows:

1. For each intersection cluster, determine the number of connected non-intersection clusters (i.e., $m$) and the dimension of each non-intersection cluster (i.e., $d_1, \cdots, d_m$);

2. Suppose there are $n$ different sorted numbers in $\{d_1, \cdots, d_m\}$, i.e., $d^1 < \cdots < d^n$. Assign the possible number of clusters to the range from $n$ to $m$. For each specified number, the dimension for each cluster is given by one number in $\{d^1, \cdots, d^n\}$ starting from the lowest to the highest, and at least one cluster has dimension $d^j, j = 1, \cdots, n$.

3. For each given number and dimensions of the clusters, compute its ED if the reconstruction error by $K$-planes is smaller than a specified maximum error $\zeta_{\max}$. Otherwise, ED is set to be the maximum number $N_{\max} = 100$.

4. The best number of clusters and their dimensions are given by the structure with the minimum ED.

Our proposed mumCluster reveals that there are three intersection clusters for the points sampled from Figure 1, where each cluster is connected with $m = 4$ non-intersection clusters. The possible structure (in the form of $(d_1^*, d_2^*, \cdots, d_{k^*}^*)$ for $k^*$ clusters) and their corresponding effective dimension are tabulated in Table 1.

Figure 2 shows the Pseudo-code of mumCluster.

**Table 1.** Effective dimension (ED) for each intersection cluster in Figure 1 w.r.t the possible structure (the best is marked in boldface).

| STRUCTURE | 2 | (2,2) | (2,2,2) | (2,2,2,2) |
|---|---|---|---|---|
| INTERSECTION CLUSTER 1 | 100 | **2.021** | 2.031 | 2.041 |
| INTERSECTION CLUSTER 2 | 100 | **2.019** | 2.029 | 2.039 |
| INTERSECTION CLUSTER 3 | 100 | **2.020** | 2.030 | 2.040 |

### 3.3 Complexity Analysis

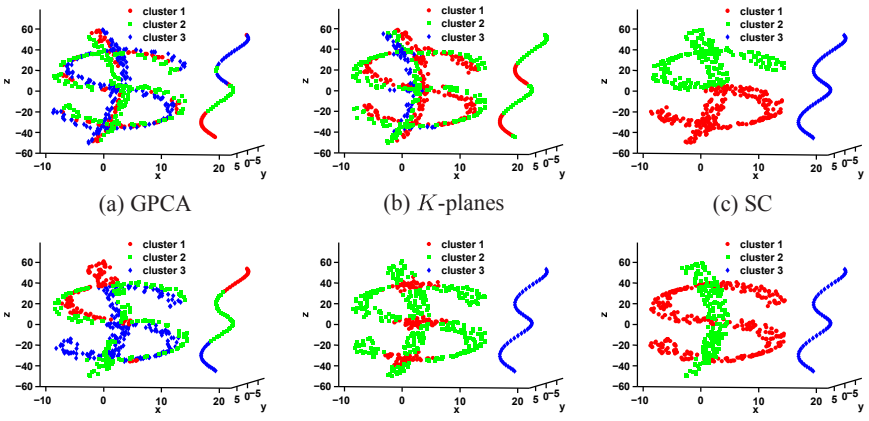
The computational complexity of our proposed mumCluster is dominated by three parts: intrinsic dimension estimation, connected components search and fine clusters identification. Intrinsic dimensions of $N$ $D$-dimensional data points are estimated by performing local PCA on $L$ nearest neighbors of each point, the complexity is $N \times O(LD\min(L,D))$. Spectral clustering is used to search for the $r$ connected components, with the total complexity $O((D+L+r)N^2 + Nr^2t)$, where $O((D+L)N^2)$ stands for the time complexity of constructing similarity graph, $O(rN^2)$ stands for the complexity of computing the first $r$ generalized eigenvectors and $O(Nr^2t)$ is the complexity of $K$-means in $r$-dimensional space for $t$ iterations. Since $r \ll N$, $L \ll N$ and $K$-means converges very quickly, the complexity of connected components search is limited by $O(N^2 \max(D,N))$. The complexity analysis of grouping fine clusters using $K$-planes is not straightforward, since we do not know the exact number of points to be grouped and a bottom-up scheme as shown in Subsection 3.2 is needed to automatically determine the number of clusters and their dimensions. However, following the same analysis in [8], the overall worst case time complexity (an upper bound) of this procedure is $O(m^2) \cdot O(DN\min(D,N))$ when there are $m$ non-intersection clusters. Note that, this result does not reflect its real running time as demonstrated by the experiments presented in the next section. To sum up, the computational complexity of mumCluster is limited by $O(N^2 \max(D,N))$ in total, which is determined by the number of data points and the number of features.

## 4 Experiments

We now evaluate the performance of our mumCluster using synthetic data and real data. Note that the number of manifold clusters and their dimensions are provided for all the other manifold clustering methods except for mumCluster. For spectral clustering (SC), the unsymmetrical normalized spectral clustering [10] is used.

### 4.1 Hybrid Modeling Data

The hybrid modeling data shown in Figure 1 are drawn from one helix, one swiss-roll, and one two-dimensional surface in $\Re^3$. The number of points are 200, 1000, 600, respectively. As we can see from Figure 3, all the other methods do not work well on this data set. Table 2 reports the clustering accuracy of the different methods. Obviously, our method performs quite well. GPCA and $K$-planes do not work well in this nonlinear




(a) GPCA (b) $K$-planes (c) SC

(d) $K$-manifolds (e) Cao-Haralick (f) MumCluster

**Fig. 3.** Grouping results using different manifold clustering methods.

**Table 2.** Clustering accuracy (%) of the different methods on the hybrid modeling data.

| GPCA | $K$-PLANES | SC | $K$-MANIFOLDS | CAO-HARALICK | MUMCLUSTER |
|---|---|---|---|---|---|
| 38.11 | 40.06 | 57.39 | 40.39 | 60.17 | **99.06** |

case because of their linear nature, while the method of Cao and Haralick treats the intersections as clusters. SC diffuses wrong clustering information across the intersecting manifolds, while $K$-manifolds fails to estimate faithful geodesic distances when there are separated clusters.

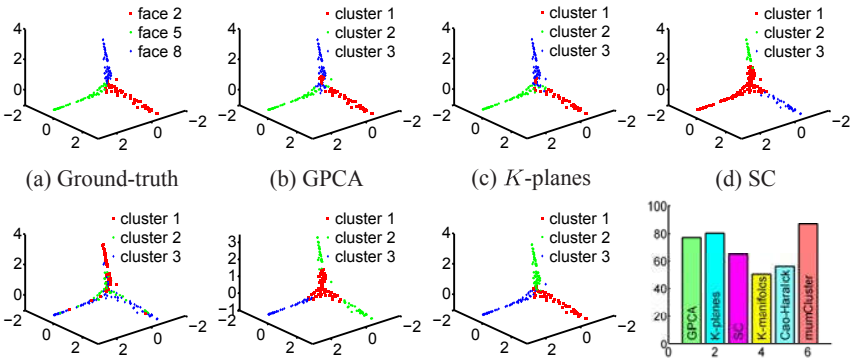
### 4.2 Single Modeling Data

It is interesting to compare our mumCluster with SC on data containing multiple single manifolds, and compare with $K$-manifolds on data containing intersecting manifolds, where SC and $K$-manifolds can work well, respectively. It is easy to see that when points are sampled from multiple separated single manifolds, our mumCluster is in fact as same as SC and therefore the results are not presented here due to the space limit. In the following, we compare mumCluster with $K$-manifolds on data containing intersecting manifolds. The spirals data set[1] (see Figure 1 of [4]) where $K$-manifolds can work well is used for the comparison. We run mumCluster and $K$-manifolds over five random samplings from this evaluated data set, as well as the other methods which can be used for intersecting manifolds. Table 3 reports the clustering accuracy. The results demonstrate that mumCluster generally outperforms the other methods.

[1] http://www.cs.wustl.edu/ rms2/kmanifolds.htm.

**Table 3.** Clustering accuracy (%) over five random samplings from the spirals data set.

| DATA SET | A | B | C | D | E |
|---|---|---|---|---|---|
| GPCA | 48.8 | 42.4 | 43.6 | 44.8 | 47.0 |
| $K$-PLANES | 48.2 | 40.6 | 49.4 | 46.4 | 46.4 |
| CAO-HARALICK | 52.0 | 50.6 | 47.6 | 51.0 | 48.4 |
| $K$-MANIFOLDS | 98.0 | 96.0 | 97.6 | 97.6 | 96.6 |
| MUMCLUSTER | **100.0** | **99.8** | **100.0** | **99.6** | **99.2** |




(a) Ground-truth (b) GPCA (c) $K$-planes (d) SC

(e) $K$-manifolds (f) Cao-Haralick (g) MumCluster (h) Accuracy (%)

**Fig. 4.** Clustering results using different methods on a subset of the Yale Face Database B.

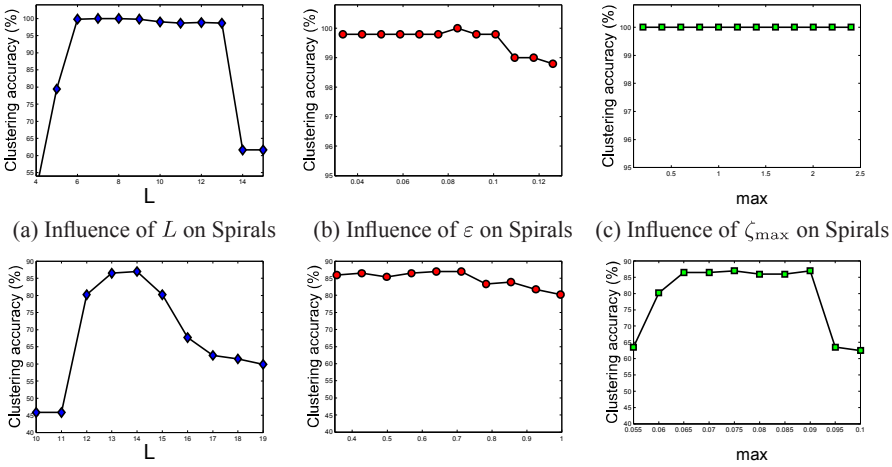
### 4.3 Illumination Variant Face Clustering

In this experiment, the face images in the Yale Face Database B[2] [18] under 64 varying lighting condition are used. We strictly follow the experimental design of [5] for a fair comparison, that is, subjects 2, 5, and 8 of this database are used and the original data are projected onto low-dimensional space (here, LLE [3] method is adopted ) before manifold clustering. For the purpose of visualization, we use the class information to label the sample as shown in Figure 4 (a), which will be used as the ground-truth for comparing the different approaches. Note that the class information of the samples are not provided to the clustering methods. We apply mumCluster and the other methods to group the data. As can be seen from Figure 4, our proposed method achieves a better clustering, which has a clustering accuracy of 86.98%, while the clustering accuracy of the other methods are 77.08%, 80.21%, 65.10%, 51.04%, 56.25%, respectively. The total running time of mumCluster on this real-world data is 0.64s, where local intrinsic dimension estimation costs 0.07s while fine clusters identification costs 0.31s.

### 4.4 The Influence of Parameters

There are three parameters in mumCluster. In this subsection, we examine their impact on the performance of mumCluster by fixing two parameters and varying the concerned

[2] http://www.cs.uiuc.edu/homes/dengcai2/Data/FaceData.html




(a) Influence of $L$ on Spirals (b) Influence of $\varepsilon$ on Spirals (c) Influence of $\zeta_{\max}$ on Spirals

(d) Influence of $L$ on Yale (e) Influence of $\varepsilon$ on Yale (f) Influence of $\zeta_{\max}$ on Yale

**Fig. 5.** Influence of parameters on mumCluster.

parameter. The results on the spirals data set A and the Yale Face Database B are plotted in Figure 5. We have studied on many other data sets, and the results are similar and thus omitted due to page limit. In general, the optimal values of these parameters depend on the distribution of the samples, while it is easy to see that mumCluster can achieve good performance over a broad rang of these parameters. In detail, the performance of mumCluster is generally insensitive to the setting of $L$, as long as it is neither too small nor too large. The reason is that $L$ is the number of nearest neighbors which will not capture enough structure information and may lead to many disconnected subgraphs when it is too small, while local property will lose when it is too large. Moreover, as we can see that the results on the Yale data have more fluctuation than on the synthetic data, which show the complexity of the real-world data and thus more attention should be paid to parameter setting. The performance of mumCluster will degenerate when $\varepsilon$ is large. The reason is that $\varepsilon$ controls the enlarged area of the intersection points, and it will become too large to ensure a locally linear area. MumCluster is relatively insensitive to the setting of $\zeta_{\max}$ , as we can see in Figure 5 (c) and (f).

Overall, Figure 5 shows that setting the parameters of mumCluster is not difficult, since the performance of mumCluster is robust to a broad range of parameter values. Moreover, among the three parameters, $L$ has more influence on the performance of mumCluster, which shows that local intrinsic dimension estimation is a key step in our scheme. However, more sophisticated intrinsic dimension estimator can be incorporated into mumCluster to improve the performance, which is our ongoing work.

## 5 Conclusion

In this paper, we propose a new manifold clustering method, i.e., mumCluster, which can work well when the samples are drawn from hybrid modeling and can adaptively

determine the number of clusters and the intrinsic dimensions. Experimental results show that mumCluster is superior to many state-of-the-art manifold clustering methods.

**Acknowledgments.** The authors are grateful to the referees for their helpful comments. This work was done when Y. Wang was visiting the LAMDA Group, Nanjing University. This work was partially supported by the NSFC (60975038, 60975043), 973 Program (2010CB327903), JiangsuSF (BK2008018) and Jiangsu 333 Program.

## References

1. Hartigan, J. A., Wong, M. A.: A K-Means Clustering Algorithm. Applied Statistics 28, 100–108 (1979)
2. Seung, H. S., Lee, D. D.: Cognition - the Manifold Ways of Perception. Science 290(5500), 2268–2269 (2000)
3. Roweis, S. T., Saul, L. K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science 290(5500), 2323–2326 (2000)
4. Souvenir, R., Pless, R.: Manifold Clustering. In: the Tenth IEEE International Conference on Computer Vision, pp. 648–653 (2005)
5. Vidal, R., Ma, Y., Sastry, S.: Generalized Principal Component Analysis (GPCA). IEEE Trans. Pattern Anal. Mach. Intell. 27(12), 1945–1959 (2005)
6. Vidal, R., Tron, R., Hartley, R.: Multiframe Motion Segmentation with Missing Data using Powerfactorization and GPCA. International Journal on Computer Vision 79(1), 85–105 (2008)
7. Bradley, P. S., Mangasarian, O. L.: K-plane Clustering. Journal of Global Optimization 16(1), 23–32 (2000)
8. Cappelli, R., Maio, D., Maltoni, D.: Multispace KL for Pattern Representation and Classification. IEEE Trans. Pattern Anal. Mach. Intell. 23(9), 977–996 (2001)
9. Haralick, R., Harpaz, R.: Linear Manifold Clustering in High Dimensional Spaces by Stochastic Search. Pattern Recognition 40(10), 2672–2684 (2007)
10. Shi, J. B., Malik, J.: Normalized Cuts and Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22(8), 888–905 (2000)
11. Ng, A., Jordan, M., Weiss, Y.: On Spectral Clustering: Analysis and an Algorithm. In: Advances in Neural Information Processing Systems 14, pp. 849–856 (2001)
12. Cao, W. B., Haralick, R.: Nonlinear Manifold Clustering by Dimensionality. In: the 18th International Conference on Pattern Recognition, pp. 920–924 (2006)
13. Hastie, T., Tibshirani, R., Friedman, J.: Elements of Statistical Learning. Springer Verlag (2001)
14. von Luxburg, U.: A Tutorial on Spectral Clustering. Statistics and Computing 17(4), 395–416 (2007)
15. Goldberg, A., Zhu, X., Singh, A., Xu, Z., Nowak, R.: Multi-manifold Semi-supervised Learning. In: the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 169–176 (2009)
16. Fukunaga, K., Olsen, D. R.: Algorithm for Finding Intrinsic Dimensionality of Data. IEEE Transactions on Computers c-20(2), 176–183 (1971)
17. Huang, K., Ma, Y., Vidal, R.: Minimum Effective Dimension for Mixtures of Subspaces: a Robust GPCA Algorithm and its Applications. In: the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 631–638 (2004)
18. Georghiades, A., Belhumeur, P., Kriegman, D.: From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose. IEEE Trans. Pattern Anal. Mach. Intell. 23(6), 643–660 (2001)