Local and Structural Consistency for Multi-manifold Clustering*

Yong Wang[1,2], **Yuan Jiang**[2], **Yi Wu**[1], **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

Data sets containing multi-manifold structures are ubiquitous in real-world tasks, and effective grouping of such data is an important yet challenging problem. Though there were many studies on this problem, it is not clear on how to design principled methods for the grouping of multiple hybrid manifolds. In this paper, we show that spectral methods are potentially helpful for hybrid manifold clustering when the neighborhood graph is constructed to connect the neighboring samples from the same manifold. However, traditional algorithms which identify neighbors according to Euclidean distance will easily connect samples belonging to different manifolds. To handle this drawback, we propose a new criterion, i.e., local and structural consistency criterion, which considers the neighboring information as well as the structural information implied by the samples. Based on this criterion, we develop a simple yet effective algorithm, named Local and Structural Consistency (LSC), for clustering with multiple hybrid manifolds. Experiments show that LSC achieves promising performance.

1 Introduction

Data sets containing low-dimensional multi-manifold structures are ubiquitous in real-world tasks, and effective grouping of such data is an important yet challenging problem, which is referred as *manifold clustering* [Souvenir and Pless, 2005; Wang *et al.*, 2010]. Many important problems of practical interest can be properly formalized under this framework. For instance, motion segmentation problem in computer vision, the point correspondences in a dynamic scene generally contain several moving objects each can be represented as a manifold; In classification of face images, the faces of the same person lie on the same manifold while different persons are associated with different manifolds.

Traditional clustering algorithms (e.g., K-means) have difficulty in effectively grouping multi-manifold data, which has attracted numerous researchers to study on manifold clustering during the last decade. Whereafter, several algorithms have been proposed, which can be roughly classified into two categories, i.e., linear and nonlinear. However, all linear algorithms fail to deliver good performance in the presence of nonlinear structures due to their linear nature. Thus, more attention should be paid to *nonlinear manifold clustering*, since a large number of recent research efforts have shown that the distributions of many natural data are inherently nonlinear [Saul and Roweis, 2004]. Existing nonlinear algorithms focus on dealing with well-separated manifolds or intersecting manifolds. However, a more accurate and general description of the real-life data is that samples are supported on a mixture of hybrid manifolds. That is, the manifolds on which the data points lie are (a) linear and/or nonlinear and (b) intersecting and/or not intersecting, which is referred as *hybrid nonlinear manifold clustering* [Wang *et al.*, 2010]. Existing algorithms either could not effectively deal with such situation, or are quite heuristic. Therefore, an open and challenging problem in manifold clustering is how to design more principled methods for the grouping of multiple hybrid manifolds.

Inspired by the practical successes and rapidly developing theoretical underpinnings of spectral clustering algorithms, in this paper, we try to design a new spectral-based algorithm to the task of detecting multiple low-dimensional manifolds embedded in high-dimensional point clouds. Firstly, we study the potential of spectral methods for hybrid nonlinear manifold clustering and our analysis reveals that they have the potential only when the neighborhood graph is constructed to connect the neighboring samples from the same manifold. However, traditional spectral clustering algorithms [Shi and Malik, 2000; Ng *et al.*, 2001], which identify neighbors according to Euclidean distance, are not suitable since samples belonging to different manifolds may be close to each other, especially near the intersection of different manifolds. To handle this drawback, our basic idea is to take advantage of additional structural information to identify the neighbors of each point from the same manifold rather than the entire Euclidean space. In doing that, we propose local and structural consistency criterion, which considers the neighboring information as well as the structural information implied by the samples. Based on this new criterion, we develop a simple yet effective spectral-based algorithm, named Local and Structural Consistency (LSC), for the task of detection of multiple

*This research was partially supported by NSFC (60975038, 61005003, 61073097) and 973 Program (2010CB327903)

hybrid manifolds. Experimental results on both synthetic and real-world data show its superior performance over existing manifold clustering algorithms.

The novelty of this work lies in: 1) the proposal of combining local consistency with structural consistency to address the fundamental problem of neighborhood selection for manifold clustering; 2) the method for constructing the local tangent space when there are intersecting clusters. To the best of our knowledge, we propose the first solutions to these two challenging problems.

The rest of this paper is organized as follows. Section 2 gives a brief review of the related works. Section 3 studies the potential of spectral methods for hybrid nonlinear manifold clustering and presents the LSC algorithm. Section 4 reports experiments on both synthetic and real-world data. Finally, we conclude in Section 5.

2 Related Work

Unlike classical clustering algorithms which are based on the concept that a cluster is centered around a single point, manifold clustering algorithms regard a cluster as a group of points compact around a low-dimensional manifold [Souvenir and Pless, 2005; Wang *et al.*, 2010].

When only a mixture of linear or affine subspaces are used to represent the clusters, the corresponding problem, i.e., *linear manifold clustering*, is much easier to address since there are elegant representations for linear manifolds. Several linear manifold clustering algorithms have been proposed in the past decade and successfully applied to many real-life situations. Generalized Principal Component Analysis (GPCA) [Vidal *et al.*, 2005] and K-planes [Tseng, 2000] learn a series of offset vectors and basis vectors from the data to represent the linear manifolds and then assign each point to the nearest linear manifold. Local Subspace Affinity (LSA) [Yan and Pollefeys, 2006] constructs an affinity matrix for any pair of points based on principal subspace angles of the local estimated tangent spaces to group the data. Spectral Curvature Clustering (SCC) [Chen and Lerman, 2009] addresses linear manifold clustering by building a polar curvature affinity matrix among certain fixed-size subset of the points to measure their possibilities of coming from the same manifold, and then uses multi-way spectral methods to group the samples.

Unlike its simple linear counterpart, nonlinear manifold clustering is more complex and difficult. The research on this topic is divided into three branches: intersecting, separated and hybrid. K-manifolds [Souvenir and Pless, 2005] which begins by estimating geodesic distances between pairwise points and then employs a EM-like algorithm to cluster samples in terms of geodesic distances, is primarily motivated for dealing with unorganized data nearly lying on multiple intersecting low-dimensional manifolds. When the observations are lying on or close to multiple well-separated manifolds, traditional spectral clustering algorithms [Shi and Malik, 2000; Ng *et al.*, 2001] can be directly used to group the data and several studies have shown their promising results. MumCluster [Wang *et al.*, 2010] is proposed to effectively deal with the general hybrid nonlinear manifold clustering problem where some data manifolds are separated but some are intersected. However, it is quite heuristic and heavily relies on the correct local dimension estimation. Up to now, it is not clear on how to design principled methods for the grouping of multiple hybrid manifolds, which is the main focus of this paper.

3 The Proposed Approach

The main focus of this section is to study the potential of spectral methods for hybrid nonlinear manifold clustering and then to design a new spectral-based algorithm to the task of detecting multiple low-dimensional manifolds.

3.1 Preliminaries

Suppose that we are given a set of unorganized data points $X = \{x_i \in \Re^D, i = 1, \cdots, N\}$, which are sampled from $K > 1$ distinct smooth manifolds $\{\Omega_j \subseteq \Re^D, j = 1, 2, \cdots, K\}$ where each manifold has N_j points ($\sum_{j=1}^{K} N_j = N$). Moreover, some data manifolds are separated but some are intersected [Wang *et al.*, 2010]. The objective of manifold clustering is to assign each sample to the manifold it belongs to. For simplicity, we restrict the discussion to the case where all the underlying manifolds have the same dimension d ($0 < d < D$), which, together with their number K, is known.

To reveal the potential of spectral methods for hybrid nonlinear manifold clustering, we first review below the main steps of Symmetrical Normalized Spectral Clustering with L-*nearest neighbor graph* and *Simple weight* (SNSC-LS) [Ng *et al.*, 2001], which is the basis of our proposed approach:

Step 1. Constructing neighborhood graph G Put an edge between x_i and x_j if x_i is among the L-nearest neighbors of x_j according to Euclidean distance, and vice versa.

Step 2. Determining affinity matrix W If x_i and x_j are connected by G, then $w_{ij} = 1$; otherwise, $w_{ij} = 0$.

Step 3. Spectral decomposition Extract $U = [u_1, \cdots, u_K]$, the top K eigenvectors of $D^{-1/2} W D^{-1/2}$, where the degree matrix $D = diag\{W * 1\}$. Then, re-normalize each row of U to have unit length, obtaining matrix V.

Step 4. Clustering by K-means Apply K-means to the rows of V (i.e., embedded data) to find K clusters. Finally, original data points are grouped into K clusters.

3.2 Potential

Recall that the performance of clustering depends on the adopted algorithm as well as the data. When the clustering algorithm is fixed, the performance mainly relies on the data. Unlike traditional clustering where clustering algorithm (e.g., K-means) is directly performed on the data, the grouping procedure of spectral clustering is performed on the embedded data. While the embedded data are obtained by performing spectral analysis on the affinity matrix of the neighborhood graph which in turn is determined by the L-nearest neighbors of each point. This fact reveals that the performance of spectral clustering is essentially relied on how to identify neighbors. In fact, we have the following Proposition.

Proposition 1 If the L-nearest neighbors of each point are from the same manifold and L is large enough to ensure that the points from the same manifold are connected, then spectral clustering gives the perfect underlying manifolds.

Proof. Without loss of generality, we assume that a permutation of indices of the data have been done such that all the points in the k-th manifold appear before the $(k+1)$-th manifold, $k = 1, \cdots, K-1$. Since the neighbors of each point are from the same manifold, pair-wise points x_i and x_j belonging to different manifolds could not be L-nearest neighbors of each other, hence their affinity value $w_{ij} = 0$. Then, the affinity matrix W is block-diagonal with K blocks, each block corresponding to one of the manifolds. It is easy to show that $D^{-1/2}WD^{-1/2}$ is also block-diagonal which has the largest eigenvalue 1 of multiplicity K under the restriction that each manifold is connected. Moreover, the top K eigenvectors associated with eigenvalue 1 can be described as:

$$U = D^{1/2} \begin{pmatrix} 1_{N_1} & 0 & \cdots & 0 \\ 0 & 1_{N_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1_{N_K} \end{pmatrix} R^T \in \Re^{N \times K}, \quad (1)$$

where 1_{N_k} denotes the N_k-dimensional vector of ones and $R = (r_1, r_2, \cdots, r_K) \in \Re^{K \times K}$ is an orthonormal matrix. As a direct consequence of the row normalization of U, for all $k = 1, \cdots, K$ and $l = 1, \cdots, N_k$, V's rows satisfy

$$V_l^k = r_k. \quad (2)$$

In other words, the K clusters are mapped to K mutually orthonormal points in $\Re^K$. Then, the followed K-means will give grouping result corresponds exactly to the "true" clustering of the original manifolds. #

For the general case where not all the L-neighbors are from the same manifold, we have the following Theorem:

Theorem 1 [Arias-Castro, 2009] For $k = 1, \cdots, K$, let I_k denotes the index set of the k-th cluster and W_k denotes the submatrix of W corresponding to I_k. Define $\hat{D}_i^k = \sum_{j \in \mathrm{I}_k} w_{ij}$ and $\tilde{D}_i^k = \sum_{j \notin \mathrm{I}_k} w_{ij}$ as *within-cluster connection* and *between-cluster connection* of $i \in \mathrm{I}_k$, respectively. Then, under the following conditions, there are K orthonormal vectors $r_1, r_2, \cdots, r_K$ such that V's rows satisfy

$$\sum_{k=1}^{K} \sum_{i \in \mathrm{I}_k} \left\| V_i^k - r_k \right\|_2^2 \leq 8CN\delta^{-2}[K^2\varepsilon_1 + K\varepsilon_2^2]. \quad (3)$$

(A1) There exists $\delta > 0$ such that, for all $k = 1, \cdots, K$, the second largest eigenvalue of W_k is bounded above by $1 - \delta$.

(A2) There is some fixed $\varepsilon_1 > 0$, so that for all $k, l \in \{1, \cdots, K\}$ with $k \neq l$, $\sum_{i \in \mathrm{I}_k} \sum_{j \in \mathrm{I}_l} w_{ij}^2 / \hat{D}_i^k \hat{D}_j^l \leq \varepsilon_1$.

(A3) For some fixed $\varepsilon_2 > 0$, for all $k = 1, \cdots, K$ and $i \in \mathrm{I}_k$, $\tilde{D}_i^k / \hat{D}_i^k \leq \varepsilon_2 \left(\sum_{s,t \in \mathrm{I}_k} w_{st}^2 / \hat{D}_s^k \hat{D}_t^k \right)^{-1/2}$.

(A4) There is some constant $C > 0$, so that for all $k = 1, \cdots, K$ and $i, j \in \mathrm{I}_k$, $\hat{D}_i^k \ll C\hat{D}_j^k$.

From the viewpoint of neighbors, Theorem 1 suggests that when each point has more neighbors (or a higher connection) with its own group than with the others, then the embedded data (i.e., the rows of V) will form tight clusters around K well-separated points. Obviously, the followed K-means performed on the embedded data will easily give grouping result corresponds to the "true" clustering of the original manifolds.

Proposition 1 and Theorem 1 reveal the potential of spectral methods for hybrid nonlinear manifold clustering, i.e., we should try our best to identify the neighboring samples of each point from the same manifold. Moreover, they also make clear the invalidation of traditional spectral clustering algorithms to multiple intersecting or even hybrid manifolds. Traditional algorithms identify neighbors according to Euclidean distance, thus samples belonging to different manifolds will easily be connected by the neighborhood graph (especially those near the intersection of different manifolds), thus diffusing information across the wrong manifolds.

Therefore, in order to extend spectral-based algorithms to the task of detection of multiple hybrid manifolds, new and effective criteria to identify neighbors are specially desired.

3.3 Local and Structural Consistency

Our analysis in the previous subsection reveals that the key to extend spectral-based algorithms to the task of detection of multiple hybrid manifolds is trying to identify the neighbors of each point from the same manifold rather than the entire Euclidean space. To this end, our basic idea is to take advantage of the nature of manifolds to incorporate some geometric information encoded in the sampled data to identify the neighbors and then to find the correct clusters.

The type of geometric information incorporated in our proposed approach is local tangent space, based on the following observations: (a). Though the samples are from globally highly nonlinear distribution, locally, each data point and its neighbors are lying on a linear patch of the manifold [Saul and Roweis, 2004]. (b). Local tangent space at each point provides a reasonable approximation to the local geometric structure of the nonlinear manifold [Zhang and Zha, 2005]. (c). Near the intersection of different manifolds, the points from the same manifold have similar local tangent spaces (i.e., they are *structural consistency*) while the points belonging to different manifolds have dissimilar tangent spaces.

However, it should be noted that only structural consistency (i.e., neighbors have similar local tangent spaces) is not enough to be used to determine neighbors. One counterexample is that the points on the right-and-left corners of S-curve in Figure 2 (a) have similar local tangent spaces to the points on the vertical affine plane yet they are from different manifolds. This fact suggests that neighbors should not only be structural consistency, but also should be *local consistency*, that is they should be close to each other.

In doing that, we have a simple criterion (named as *local and structural consistency criterion*) to identify neighbors, which selects L neighbors of each point using the following measure of distance:

$$Dis(x_i, x_j) = \|x_i - x_j\|_2^2 + \lambda * \|\Theta_i - \Theta_j\|_F^2, \quad (4)$$

where $Dis(x_i, x_j)$ is our measure of distance, $\|x_i - x_j\|_2$ is the Euclidean distance between x_i and x_j, and $\|\Theta_i - \Theta_j\|_F$ is the *projection F-norm distance* between the local tangent spaces of x_i and x_j (denoted as Θ_i and Θ_j, respectively). In (4), λ is a parameter which balances the contribution of local consistency and structural consistency.

After the L-nearest neighbors have been identified by our new measure of distance, the following procedures to group clusters are the same as traditional spectral clustering algorithms, thus our method to the task of detection of multiple hybrid manifolds is named as Local and Structural Con-

sistency (abbreviated as LSC). The pseudo-code of LSC is shown in Figure 1.

3.4 Implementation

From (4) and Figure 1, we can see that the key to implement local and structural consistency criterion and the LSC algorithm are boiled down to effectively approximate the local tangent space at each sample, which will be discussed below.

Our basic idea to approximate the local tangent spaces is based on the following facts: (a). Globally nonlinear manifolds can locally be well-approximated by a series of local linear manifolds [Saul and Roweis, 2004]. (b). Probabilistic principal component analyzers [Tipping and Bishop, 1999] can successfully pass across the intersecting linear manifolds. (c). The points approximated by the same probabilistic analyzer usually have similar local tangent spaces which can be well-approximated by the principal subspace of the corresponding analyzer. Thus, we can train a set of probabilistic analyzers to approximate the nonlinear manifolds, and then determine the local tangent space of a given sample as the principal subspace of its corresponding analyzer.

Specifically, we train M mixtures of probabilistic principal component analyzers [Tipping and Bishop, 1999], each of which is characterized by the model parameters $\theta_m = \{\mu_m, V_m, \sigma_m^2\}$, $m = 1, \cdots, M$, where $\mu_m \in \Re^D$ is the mean vector, $V_m \in \Re^{D\times d}$ and σ_m^2 is a scalar. Under the m-th analyzer, observed data x is related to a corresponding latent variable $y \sim \mathcal{N}(0, I)$ and noise $\varepsilon_m \sim \mathcal{N}(0, \sigma_m^2 I)$ as:

$$x = V_m y + \mu_m + \varepsilon_m. \tag{5}$$

Then, the marginal distribution of x is given in the form

$$p(x\,|m\,) = (2\pi)^{-D/2}|C_m|^{-1/2}\exp\left\{-\tfrac{1}{2}(x-\mu_m)^{\mathrm{T}}C_m^{-1}(x-\mu_m)\right\}, \tag{6}$$

where $C_m = \sigma_m^2 I + V_m V_m^T$. Moreover, the log-likelihood of observing the data set X for M analyzers is:

$$\mathcal{L} = \sum\nolimits_{i=1}^{N} \ln\left\{\sum\nolimits_{m=1}^{M} \pi_m p(x_i\,|m\,)\right\}, \tag{7}$$

where $\pi_m \geq 0$ is the mixing proportion with $\sum_{m=1}^{M}\pi_m = 1$.

We can learn all the model parameters μ_m, V_m, and σ_m^2 by developing an iterative EM algorithm (initialized with K-means) to maximize (7). Due to page limitation, the principal steps of this EM learning procedure are omitted, more details can be found in [Tipping and Bishop, 1999].

Then, we group sample x_i into the j-th probabilistic analyzer, which maximizes the posterior responsibility of x_i:

$$p(j\,|x_i\,) = \max_m \;\; p(m\,|x_i\,), \tag{8}$$

where the posterior responsibility $p(m\,|x_i\,)$ is given by

$$p(m\,|x_i\,) = p(x_i\,|m)\pi_m \Big/ \sum\nolimits_{m=1}^{M} p(x_i\,|m)\pi_m. \tag{9}$$

Finally, the local tangent space of x_i is approximated by

$$\Theta_i = span(V_j). \tag{10}$$

The reconstruction error of using M probabilistic analyzers to approximate the underlying manifolds is:

$$error(M) = \sum\nolimits_{j=1}^{M}\sum\nolimits_{l=1}^{m_j}(x_l^j - \mu_j)^{\mathrm{T}}(I - V_j V_j^{\mathrm{T}})(x_l^j - \mu_j), \tag{11}$$

where $x_l^j, l = 1, \cdots, m_j$ are the m_j ($\sum_{j=1}^{M} m_j = N$) points which are grouped into the j-th probabilistic analyzer.

Algorithm Local and Structural Consistency (LSC)

Input: Data set X, number of clusters K, dimension of the manifolds d, number of mixture models M, number of neighbors L, balance parameter λ.

Steps: 1: Compute the local tangent space of each point.
2: Identify L-nearest neighbors of each point according to (4).
3-5: Same as steps 2-4 of the SNSC-LS algorithm.

Output: A partition of the data into K clusters.

Figure 1: Pseudo-code of the LSC algorithm

3.5 Complexity Analysis

Since the computational complexity of a clustering algorithm is much important to its applicability, let us briefly analyze the time complexity of the LSC algorithm.

The computational complexity of estimating the local tangent space of each point is $O(NDM(t_1 + dt_2))$, where t_1 and t_2 are the number of iterations before K-means and EM to convergence, respectively. Moreover, the complexity of computing the projection F-norm distances between any two local tangent spaces are $O(M^2Dd^2)$. Steps 2-5 of LSC have the same time complexity as SNSC-LS, which is $O(N^3 + N^2D + NK^2t_3)$ where t_3 is the number of iterations before K-means to convergence in step 5. Therefore, the total time complexity of the LSC algorithm is

$$O(N^3 + N^2D + N(DM(t_1 + dt_2) + K^2t_3) + M^2Dd^2). \tag{12}$$

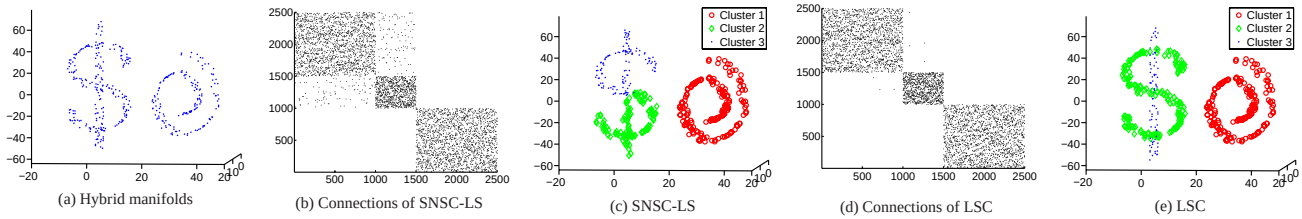
4 Experimental Results

In this section, we investigate the performance of LSC using a series of synthetic and real-world data sets. We also compare it with several state-of-the-art algorithms. All of our experiments are performed on a PC machine configured with Intel Dual Core based system with $4 * 2.6$ GHz CPU and 8GB RAM memory under Matlab platform.

4.1 Visual Comparison

We first visually compare LSC with SNSC-LS to the task of detection of multiple hybrid manifolds. The data points, as shown in Figure 2 (a), are sampled from three synthetically generated hybrid nonlinear manifolds: one S-curve, one two-dimensional affine plane, and one Swiss-roll in $\Re^3$. The numbers of points are 1000, 500 and 1000, respectively. The connections among different points, as well as the grouping results of our LSC and SNSC-LS are shown in the last four sub-figures of Figure 2. Note that, to help the visualization of the connections, we have ordered the points so that all points belonging to the first manifold appear firstly and the points in the second manifold secondly.

As can be seen from Figure 2 (d) and (e), the performance of the proposed LSC which uses the new measure of distance (4) to select neighbors is surprisingly good: LSC identifies



(a) Hybrid manifolds (b) Connections of SNSC-LS (c) SNSC-LS (d) Connections of LSC (e) LSC

Figure 2: Grouping results of different methods on multiple hybrid manifolds: (a) Data points sampled from multiple hybrid manifolds where some manifolds are separated but some are intersected. (b) Connection relationships of SNSC-LS. (c) Grouping result of SNSC-LS. (d) Connection relationships of LSC. (e) Grouping result of LSC.

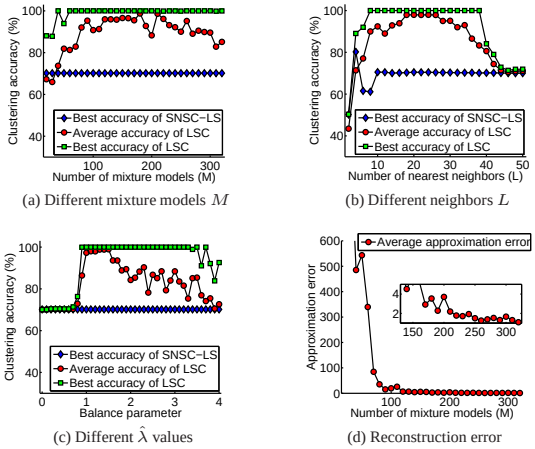
neighbors of each point generally from the same manifold and reliably finds clusters consistent with different manifolds. While SNSC-LS which selects neighbors according to Euclidean distance has many neighbors from different manifolds (Figure 2 (b)) and thus confuses points from the two intersecting manifolds though it can correctly partition the separated Swiss-roll (Figure 2 (c)).

These visual experiments validate our study of the potential of spectral methods for hybrid nonlinear manifold clustering and the effectiveness of the proposed LSC algorithm.

4.2 Parameter Influence

There are three adjustable parameters in LSC, i.e., M, L, and λ. In this subsection, we examine their influence to the performance of LSC and then try to give some default values to their setting. The results performed on the hybrid nonlinear manifolds of Figure 2 (a) over 10 independent trials are plotted in Figure 3. Note that, we set $\lambda = \hat{\lambda} * mLnn(X)$, where $mLnn(X)$ is the mean squared Euclidean distance of all samples's L-th nearest neighbors in X and $\hat{\lambda}$ is a scalar. This setting is based on two concerns: 1) to balance the magnitude between Euclidean distance and projection F-norm distance; 2) the most importantly, as revealed by our analysis, the key to the success of manifold clustering is to identify the neighbors of each point from the same manifold/cluster, and thus the added projection F-norm distance should ensure that the smallest "distance" between two inter-cluster points is larger than the "distance" between any point and its L-th intra-cluster nearest neighbors, which motivates us to bring $mLnn(X)$ into the setting of λ.

As can be seen from Figure 3, LSC achieves significantly better results than traditional spectral clustering over a large range of parameter setting. In detail, we have the following observations: (a). LSC works well when the number of mixture models M is large enough (see Figure 3 (a)), which can be explained from Figure 3 (d): as the number of mixture models increasing, the average approximation error decreasing which means that the approximation to the local linear patches of the manifolds and the local tangent space of each sample become more faithful. Therefore, better clustering accuracy is given since LSC relies on the correct estimation of these tangent spaces. (b). The performance of LSC is robust for a range of values of L, as long as it is neither too small nor too large. This phenomenon is consistent with many existing observations [Saul and Roweis, 2004] since there may be many disconnected sub-clusters when L is too small, while local consistency will lose when it is too large. (c). LSC works well when $\hat{\lambda}$ is relatively large but not too large. The reason is that the larger $\hat{\lambda}$, the more structural consistency is



(a) Different mixture models M (b) Different neighbors L

(c) Different $\hat{\lambda}$ values (d) Reconstruction error

Figure 3: Influence of parameters for LSC.

incorporated. However, local consistency becomes relatively small as $\hat{\lambda}$ is too large. This fact confirms the validity of balancing between local consistency and structural consistency.

Finally, we can give some recommended values to the setup of these parameters which will be generally used to give the reported results. As a general recommendation we suggest to work with $M = \lceil N/(7d) \rceil$, $L = 2\lceil \log(N) \rceil$ and $\hat{\lambda} = 1.2$. We can work with a relative small M, e.g., $M = 3K$, when all the manifolds are linear.

4.3 Comparison on Benchmark Data Sets

In this subsection, we compare LSC algorithm with a considerable amount of state-of-the-art algorithms in manifold clustering: GPCA [Vidal *et al.*, 2005], K-planes [Tseng, 2000], LSA [Yan and Pollefeys, 2006], SCC [Chen and Lerman, 2009], K-manifolds [Souvenir and Pless, 2005], SNSC-LS [Ng *et al.*, 2001] and mumCluster [Wang *et al.*, 2010]. We compare all these algorithms on four benchmark data sets with a wide variety of structures which cover all aspects of manifold clustering: three linear planes [Vidal *et al.*, 2005] in $\Re^3$, two well-separated circles [Ng *et al.*, 2001] and two intersecting spirals [Souvenir and Pless, 2005] in $\Re^2$, and the hybrid data as shown in Figure 2 (a).

The average clustering accuracy with the corresponding standard deviation and the average computation time over 10 independent runs are tabulated in Table 1. The best average accuracy are boldfaced. The results reveal several interesting observations: (a). Different from many state-of-the-art algorithms which are designed for a specific manifold clustering problem, such as linear algorithms (i.e., GPCA, K-planes, LSA and SCC) for linear manifold clustering and K-manifolds for intersecting manifolds, both mumCluster and

Table 1: Comparison of the clustering accuracy (mean±std.) and the average computation time (in seconds) of different algorithms on four synthetic data sets with their characteristics (N, D, d) in the parentheses.

Data set	Three-planes (1200, 3, 2)		Two-circles (600, 2, 1)		Two-spirals (1000, 2, 1)		Hybrid data (2500, 3, 2)	
	accuracy	time	accuracy	time	accuracy	time	accuracy	time
GPCA	0.983±0.000	0.01	0.502±0.000	0.01	0.543±0.000	0.01	0.458±0.000	0.02
K-planes	0.944±0.124	0.01	0.502±0.002	0.01	0.538±0.001	0.01	0.351±0.008	0.05
LSA	0.976±0.000	56.03	0.502±0.000	11.60	0.519±0.000	35.03	0.569±0.000	304.74
SCC	0.984±0.001	1.41	0.502±0.001	0.42	0.516±0.014	0.60	0.570±0.085	3.49
K-manifolds	0.793±0.164	511.04	0.529±0.028	49.55	0.701±0.226	90.84	0.417±0.017	3670.26
SNSC-LS	0.405±0.001	1.37	**1.000±0.000**	0.23	0.708±0.003	1.29	0.702±0.000	8.57
mumCluster	0.984±0.000	6.13	**1.000±0.000**	1.31	0.859±0.004	4.74	0.984±0.000	34.09
LSC	**0.985±0.002**	3.51	**1.000±0.000**	2.24	**0.996±0.003**	6.09	**0.987±0.040**	28.96

the proposed LSC give promising performance to all cases of manifold clustering. (b). The proposed LSC performs better than SNSC-LS when there are intersecting manifolds in the data and it always achieves the best results. (c). Generally, as we have expected, the running time of LSC is comparable to many of other algorithms.

4.4 Experiments on Real Data

In this subsection, we compare the performance of LSC with the state-of-the-art algorithms on a well-known yet challenging object recognition database, i.e., the processed COIL-20 database [Nene *et al.*, 1996]. As illustrated in Figure 4, the objects have a wide variety of complex geometric, appearance and reflectance characteristics. Several studies have shown that this task can be viewed as a manifold clustering problem [van der Maaten and Hinton, 2008].

Figure 4: Twenty images from the COIL-20 database.

The whole database ($N = 1440$) and three objects subject to the same topic ($N = 216$), i.e., three different cars, are used in our experiments. The original resolution of these images is 32×32. Here, for computational efficiency, all the algorithms are executed in 10-dimensional principal components space, except that GPCA is executed in 5-dimensional space since it requires a low-dimensional space to work properly [Vidal *et al.*, 2005]. The results for all the algorithms are shown in Table 2, which show that LSC is superior to other algorithms, especially on the coil-three-cars data set.

Table 2: Comparison of the clustering accuracy (mean±std.) and the average computation time (in seconds).

Data set	COIL-20		COIL-three-cars	
	accuracy	time	accuracy	time
GPCA	0.311±0.000	159.45	0.347±0.000	0.01
K-planes	0.515±0.045	0.27	0.345±0.004	0.01
LSA	0.630±0.029	79.93	0.356±0.000	1.39
SCC	0.647±0.035	27.62	0.351±0.005	0.78
SNSC-LS	0.717±0.031	3.34	0.356±0.002	0.09
mumCluster	0.621±0.053	18.79	0.804±0.161	0.43
LSC	**0.741±0.027**	14.46	**0.989±0.010**	0.46

5 Conclusion

In this paper, we study the potential of spectral methods for hybrid nonlinear manifold clustering. Then, we propose LSC algorithm to the task of detection of multiple hybrid manifolds. Experimental results validate its effectiveness.

References

[Arias-Castro, 2009] E. Arias-Castro. Clustering based on pairwise distances when the data is of mixed dimensions. *Available from http://arxiv.org/abs/0909.2353*, 2009.

[Chen and Lerman, 2009] G. L. Chen and G. Lerman. Spectral curvature clustering (SCC). *IJCV*, 81(3):317–330, 2009.

[Nene *et al.*, 1996] S. A. Nene, Nayar S. K., and Murase H. Columbia object image library (COIL-20). Technical Report CUCS-005-96, 1996.

[Ng *et al.*, 2001] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS 14*, pages 849–856, 2001.

[Saul and Roweis, 2004] L. K. Saul and S. T. Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *JMLR*, 4(2):119–155, 2004.

[Shi and Malik, 2000] J. B. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 22(8):888–905, 2000.

[Souvenir and Pless, 2005] R. Souvenir and R. Pless. Manifold clustering. In *ICCV*, pages 648–653, 2005.

[Tipping and Bishop, 1999] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, 1999.

[Tseng, 2000] P. Tseng. Nearest q-flat to m points. *Journal of Optimization Theory and Applications*, 105(1):249–252, 2000.

[van der Maaten and Hinton, 2008] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *JMLR*, 9:2579–2605, 2008.

[Vidal *et al.*, 2005] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). *PAMI*, 27(12):1945–1959, 2005.

[Wang *et al.*, 2010] Y. Wang, Y. Jiang, Y. Wu, and Z.-H. Zhou. Multi-manifold clustering. In *PRICAI*, pages 280–291, 2010.

[Yan and Pollefeys, 2006] J. Y. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *ECCV*, pages 94–106, 2006.

[Zhang and Zha, 2005] Z. Y. Zhang and H. Y. Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM J. Scientific Computing*, 26(1):313–338, 2005.