Intrinsic dimension estimation of data by principal component analysis

Mingyu Fan, Nannan Gu, Hong Qiao and Bo Zhang

Abstract—**Estimating intrinsic dimensionality of data is a classic problem in pattern recognition and statistics. Principal Component Analysis (PCA) is a powerful tool in discovering dimensionality of data sets with a linear structure; it, however, becomes ineffective when data have a nonlinear structure. In this paper, we propose a new PCA-based method to estimate intrinsic dimension of data with nonlinear structures. Our method works by first finding a minimal cover of the data set, then performing PCA locally on each subset in the cover and finally giving the estimation result by checking up the data variance on all small neighborhood regions. The proposed method utilizes the whole data set to estimate its intrinsic dimension and is convenient for incremental learning. In addition, our new PCA procedure can filter out noise in data and converge to a stable estimation with the neighborhood region size increasing. Experiments on synthetic and real world data sets show effectiveness of the proposed method.**

Index Terms—**Pattern recognition; Principal component analysis; Intrinsic dimensionality estimation.**

I. INTRODUCTION

Intrinsic dimensionality (ID) of data is a key priori knowledge in pattern recognition and statistics, such as time series analysis, classification and neural networks, to improve their performance. In time series analysis [1], the domain of attraction of a nonlinear dynamic system has a very complex geometric structure, and study on the geometry of the attraction domain is closely related to the fractal geometry. Fractal dimension is an important tool to characterize certain geometric properties of complex sets. In neural network design [2], the number of hidden units in the encoding middle layer should be chosen according to the ID of data. In classification tasks [3], in order to balance the generalization ability and the empirical risk value, the complexity of the function should also be related to the ID of data.

M. Fan and B. Zhang are with LSEC and the Institute of Applied Mathematics, AMSS, Chinese Academy of Sciences, Beijing 100190, China (email: fanmingyu@amss.ac.cn, b.zhang@amt.ac.cn)

N. Gu and H. Qiao are with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (email: gunannan@gmail.com, hong.qiao@ia.ac.cn)

Corresponding author: Bo Zhang

Recently, manifold learning, an important approach for nonlinear dimensionality reduction, has drawn great interests. Important manifold learning algorithms include isometric feature mapping (Isomap) [4], locally linear embedding (LLE) [5] and Laplacian eigenmaps (LE) [6]. They all assume data to distribute on an intrinsically low-dimensional sub-manifold [7] and reduce the dimensionality of data by investigating the intrinsic structure of data. However, all manifold learning algorithms require the ID of data as a key parameter for implementation.

Previous ID estimation methods can be categorized mainly into three groups: projection approach, geometric approach and probabilistic approach. The projection approach [9]–[11] finds ID by checking up the low-dimensional embedding of data. The geometric method [22] finds ID by investigating the intrinsic geometric structure of data. The probabilistic technique [19] builds estimators by making distribution assumptions on data. These approaches will be briefly introduced in Section II.

In this paper, we propose a new PCA-based method for ID estimation which is called the C-PCA method. The proposed method first finds a minimal cover of the data set, and each subset in the cover is considered as a small subregion of the data manifold. Then, on each subset, a revised PCA procedure is applied to examine the local structure. The revised PCA method can filter out noise in data and leads to a stable and convergent ID estimation with the increase of the subregion size, as shown by the experimental results. This is an advantage over the traditional PCA method which is very sensitive to noise, outliers and the choice of the subregion size. Further analysis shows that the revised PCA procedure can efficiently reduce the running time complexity and utilize all data samples for ID estimation. We should remark that our ID estimation method is also applicable to incremental learning for consecutive data. Our method is compared with the maximum likelihood estimation (MLE) method [19], the manifold adaptive method (which is referred to as the k-$k/2$ NN method in this paper) [18] and the k-nearest neighbor graph (k-NNG) method [26], [27] through experiments.

The rest of the paper is organized as follows. In

Section II, previous ID estimation methods are briefly reviewed. In Section III, the new ID estimation method (C-PCA) is introduced. In Section IV, experiments are conducted on synthetic and real world data sets to show the effectiveness of the proposed algorithm. Conclusion is made in Section V.

II. PREVIOUS ALGORITHMS ON ID ESTIMATION

Previously, there are mainly three approaches to estimate the ID of data: projection, geometric and probabilistic approaches.

The projection approach first projects data into a low-dimensional space and then determine the ID by verifying the low-dimensional representation of data. PCA is a classical projection method which finds ID by counting the number of significant eigenvalues. However, the traditional PCA only works on data lying in a linear subspace but becomes ineffective when data distribute on a nonlinear manifold. To overcome this limitation, local-PCA [9] and OTPMs PCA [10] have been proposed and can discover the ID of data lying on nonlinear manifolds by performing the PCA method locally. The Isomap algorithm yields ID of data by inspecting the elbow of residual variance curve [4]. Cheng et al. gave an efficient procedure to compute eigenvalues and eigenvectors in PCA [24].

Geometric approaches make use of the geometric structure of data to build ID estimators. Fractal-based methods have been well developed and used in time series analysis. For example, the correlation dimension (a kind of fractal dimensions) was used in [13] to estimate the ID, whilst the method of packing numbers was proposed in [14] to find the ID. Other fractal-based methods include the kernel correlation method [23] and the quantization estimator [21]. A good survey on fractal-based methods can be found in [22]. There are also many methods based on techniques from computational geometry. Lin [15] and Cheng [25] suggested to construct simplices to find the ID, while the nearest neighbor approach uses the distances between data points with their nearest neighbors to build ID estimators such as the estimator proposed by Pettis et al. [29], the k-NNG method [26], [27] and the incising ball method [17]. A comparison of the local-PCA method with that introduced by Pettis et al. was made in [16].

Probabilistic methods are based on probabilistic assumptions of data and have been tested on various data sets with stable performance. The MLE-method [19] is a representative method of this approach, whose final global estimator is given by averaging the local estimators:

$$\hat{d}_k(x_i) = \left[\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{T_k(x_i)}{T_j(x_i)} \right]^{-1}, \text{ for } i = 1, \cdots, N,$$

where $T_k(x_i)$ is the distance between x_i and its k-th nearest neighbor. MacKay and Ghahramani [20] pointed out that compared with averaging the local estimators directly, it is more sensible to average their inverses $\hat{d}_k^{-1}(x_i)$, $i = 1, \cdots, N$ for the maximum likelihood purpose. The recommended final estimator is

$$\hat{d}_k^{-1} = \frac{1}{N} \sum_{i=1}^{N} \hat{d}_k^{-1}(x_i),$$

where d_k is the estimated ID corresponding to the neighborhood size k.

III. ID ESTIMATION USING PCA WITH COVER SETS: C-PCA

Basically, there are two kinds of definitions of ID that are commonly used. One is based on the fractal dimension, such as the Hausdorff dimension and the packing dimension that are usually real positive numbers. The other kind of definition is based on the embedding manifold whose ID is always an integer.

Definition 3.1 (Embedding manifold and dimension): Let $d < D$ and let Ω be a compact open set in $\mathbb{R}^d$. Assume that span $\{\Omega - \int_\Omega d\mu\} = \mathbb{R}^d$ and $\phi : \Omega \to \mathbb{R}^D$ is a smooth function. The set $\mathcal{X} = \phi(\Omega)$ is called an embedding manifold with d its embedding dimension.

More and more real world data are proved to have nonlinear intrinsic structures and may possibly distribute on nonlinear embedding manifolds [7]. Therefore, estimation of embedding ID of data becomes an important problem [17]. In this paper, we focus on estimation of embedding dimensions.

A. PCA-based methods for ID estimation

The traditional PCA can find a subspace on which data projections have maximum variance. Given a data set $\mathcal{X} = \{x_1, \cdots, x_N\}$ with $x_i \in \mathbb{R}^D$. Let $X = [x_1, \cdots, x_N]$ and $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$. The covariance matrix of $\mathcal{X}$ is given by

$$C = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T.$$

Since C is a positive semi-definite matrix, we can assume that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$ are the eigenvalues of C with $\nu_1, \cdots, \nu_N$ the corresponding orthonormal eigenvectors, respectively. The eigen-decomposition of

matrix C is denoted as $C = \Gamma D \Gamma^T$, where D is a diagonal matrix with $D_{ii} = \lambda_i$ and $\Gamma = [\nu_1, \cdots, \nu_N]$. The eigenvector ν_i is the i-th principal direction (PD) and, for any variable x, $y_i = \nu_i x$ is defined as the i-th principal component (PC). By the definition, we have the variance $\text{var}(y_i) = \lambda_i$ and the covariance $\text{cov}(y_i, y_j) = 0$.

If the data set $\mathcal{X}$ distributes on a linear subspace, then the d primary PDs should be able to span the subspace and the corresponding PCs can account for most of the variations contained in $\mathcal{X}$. On the other hand, the variance of PCs on $\nu_{d+1}, \cdots, \nu_N$ (i.e., the PDs which are orthogonal to the linear subspace of dimension d) will be trivial. The most commonly-used criterions for ID estimation with the PCA method are

$$\frac{\min\limits_{i=1,\cdots,d}(\text{var}(y_i))}{\max\limits_{j=d+1,\cdots,N}(\text{var}(y_j))} > \alpha \gg 1 \qquad (1)$$

and the percentage of the accounted variance

$$\frac{\sum_{i=1}^{d}\text{var}(y_i)}{\sum_{i=1}^{N}\text{var}(y_i)} > \beta, \qquad 0 < \beta < 1. \qquad (2)$$

In this paper, the ID, d, is determined if the condition (1) or (2) is satisfied.

B. Filtering out the noise of data

There are two challenges for PCA-based ID estimation methods. The first one is how to filter out the noise in data, while the second one is how to choose the size of subregions on the manifold. Previously, the ID estimation of data obtained with PCA-based methods always increases with the size of subregions so the methods can not converge to give a stable ID estimation. In order to address these two limitations, we propose the following noise filtering procedure which can efficiently filter out the noise in data and make PCA-based methods to converge.

Consider the effect of additive white noise μ in data with $E(\mu) = 0$ and $\text{var}(\mu) = \sigma^2$. The covariance matrix of the noise corrupted data is given by

$$C' = \text{var}(X + \mu) = C + \sigma^2 I,$$

where C is the covariance matrix of the data $\mathcal{X}$. It can be seen that the PDs of C' are identical to those of C and the eigenvalues of C' are $\lambda_i' = \lambda_i + \sigma^2$. If σ is relatively large, then the ID criterions (1) and (2) will be ineffective.

The variance of data projections on the PDs that are orthogonal to the intrinsic embedding subspace is very small, and the most part of the variance is produced by noise. Therefore, it is possible to calculate the variance of noise by projecting data on the orthogonal PDs. Given a real number P which is very close to 1 (P is taken to be 0.95 in this paper), the noise part of data is determined by

$$\frac{\sum_{i=1}^{r-1}\text{var}(y_i)}{\sum_{i=1}^{N}\text{var}(y_i)} < P \quad \text{and} \quad \frac{\sum_{i=1}^{r}\text{var}(y_i)}{\sum_{i=1}^{N}\text{var}(y_i)} > P.$$

Thus, the variance of noise contained in data can be estimated as

$$\hat{\sigma}^2 = \frac{1}{N - r + 1}\sum_{i=r}^{N}\text{var}(y_i). \qquad (3)$$

Our new ID estimation criterions make use of the updated variance on PDs: $\text{var}(y_i) = \lambda_i - \hat{\sigma}^2$.

Remark 3.1: Noise is typically different from outliers. Noise affects every data points independently, while outliers are referred to data points that are at least at a certain distance from the data points on manifold. The proposed procedure is very robust to both noise and outliers, as shown in experiments. On the other hand, the traditional PCA procedure can handle limited noise but is very sensitive to outliers.

C. The local region selection method

An embedding manifold can be approximated locally by linear subspaces. The dimensionality of each linear subspace should be equal to the ID of the embedding manifold. Therefore, it is possible to estimate the ID of a nonlinear manifold by checking it locally. A cover is referred to a set whose elements are subsets of the data set satisfying that the union of all subsets in the cover contains the whole data set.

Definition 3.2 (The set cover problem): Given a universe $\mathcal{X}$ of N elements and a collection F of subsets of $\mathcal{X}$, where $F = \{F_1, \cdots, F_N\}$. Set cover is concerned with finding a minimum sub-collection of F that covers all data points.

Using a minimum cover has two advantages. First, it can find the minimal number of subregions, which helps save the computational time. Secondly, the result of ID estimation that utilizes the whole data set is more reliable. However, searching such a minimal cover is an NP-hard problem. In the following, we introduce an algorithm which can approximately find a minimal cover of a data set.

Given the parameter, an integer k or a real number ε, there are two ways to define the neighborhood of any data point x:

1) The k-NN method: any data point x_i that is one of k nearest data points of x is in the neighborhood of x;

2) The ε-NN method: any data point x_i in the region $\{y : \|y - x\| < \varepsilon\}$ is in the neighborhood of x.

Without loss of generality, we may assume that the index of data points is independent of their locations.

Algorithm 1 (Minimum set cover algorithm)

Input: Neighborhood size k (integer) or ε (real number), distance matrix $\hat{D} = (\|x_i - x_j\|)$
Output: Minimum cover $F = \{(F_i, r_i), i = 1, \cdots, S\}$.

1: **for** i=1 to N **do**
2:     Identify the neighbors $\{x_{i_1}, \cdots, x_{i_{P_i}}\}$ of x_i by the k-NN or ε-NN method. Let $F_i = \{i, i_1, \cdots, i_{P_i}\}$ be the index set of the neighborhood and let D be the $0-1$ incidence matrix.
3: **end for**
4: Let $F = \{(F_i, r_i = 0), i = 1, \cdots, N\}$
5: **for** $i = 1$ to N **do**
6:     Let the frequency of x_i be computed by $Q_i = \sum_{j=1}^{N} D_{ij}$.
7: **end for**
8: **for** $i = 1$ to N **do**
9:     **if** $Q_i, Q_{i_1}, \cdots, Q_{i_{P_i}} > 1$ **then**
10:         Remove (F_i, r_i) from the cover set F and set $Q_i = Q_i - 1$, $Q_{i_1} = Q_{i_1} - 1$, $\cdots$, $Q_{i_{P_i}} = Q_{i_{P_i}} - 1$.
11:     **else**
12:         Let $r_i = \max\limits_{j=1,\cdots,P_i} \|x_i - x_{i_j}\|$
13:     **end if**
14: **end for**

Using the above approximation algorithm, a cover $F = \{(F_i, r_i), i = 1, \cdots, S\}$ of the data set $\mathcal{X}$ can be found. Compared with the local region selection algorithm used in [9], our algorithm above has a low time complexity and avoids the supervised process to choose the neighborhood. Intuitively, the cardinality S of the cover F satisfies that $N/k < S < N/2k$, where k is the average number of neighbors.

D. The proposed ID estimation algorithm

We now present the proposed ID estimation algorithm using local PCA on the minimal set cover: the C-PCD algorithms, which are summarized below for both batch and incremental data, respectively.

In many cases, consecutive data are collected incrementally. This requires an incremental learning algorithm to inspect the change of the data structure on time. The incremental C-PCA algorithm is presented as follows.

Remark 3.2: Our method is different from the Local-PCA [9] in many aspects. First, the centers and the

---

Algorithm 2 (The C-PCA algorithm for batch data)

Step 1. Given a parameter k or ε, compute a minimal cover of $\mathcal{X}$ by Algorithm III-C. Without loss of generality, $F = \{(F_i, r_i) : i = 1, \cdots, S\}$ is assumed to be the constructed minimal set cover.
Step 2. Perform the PCA algorithm proposed in Subsections III-A and III-B on subsets F_i, $i = 1 \cdots, S$. The local ID estimations $\{\hat{d}_i\}_{i=1}^{S}$ are then obtained.
Step 3. Let λ_{ij} be the j-th eigenvalue on the i-th subset in the decreasing order. $\lambda_j = \sum_i \lambda_{ij}$ is considered as the variance of $\mathcal{X}$ on its j-th PD. Subsequently, the global ID estimation $\hat{d}$ can be derived using the criterions (1) or (2).

Algorithm 3 (The incremental C-PCA algorithm)

Step 1. The new data point is assumed to be x. Let $\{x_1, \cdots, x_S\}$ be the centers of the subsets in the cover. Find the nearest center x_q of x: $x_q = \arg \min\limits_{i=1,\cdots,S} \|x - x_i\|$.
Step 2. If $\|x_q - x\| > r_q$, then the data point x is considered as an outlier and the remaining part of the algorithm will not be performed on x. Otherwise, go to **Step 3**.
Step 3. Performs PCA on $F_q = F_q \bigcup \{x\}$. Let λ'_{qj} be the j-th eigenvalue. Update λ_j by $\lambda_j = \lambda_j + \lambda'_{qj} - \lambda_{qj}$. Then let $\lambda_{qj} = \lambda'_{qj}$.
Step 4. Update the local ID, $\hat{d}_q$, and the global ID, $\hat{d}$, of $\mathcal{X}$.

---

local regions are determined simultaneously by using one parameter - the neighborhood size, whilst, in [9], the centers and neighborhood sizes are determined by two parameters. Secondly, our approach finds the subregions by approximating a minimum cover of the data set, while the local-PCA in [9] does not guarantee whether or not the selected subregions cover the whole data set.

E. Computational complexity analysis

The computational complexity of our algorithms is one of the most important issues for its application. The batch mode ID estimation can be divided into two parts. In the first part, computing the distance matrix needs $O(N^2)$ time, searching the nearest neighbors for every data point needs $O(kN^2)$ time and finding an approximate minimum cover of $\mathcal{X}$ needs $O(kN)$ time. Therefore, the first part needs $O((K+1)N^2 + kN)$ running time. In the second part, performing PCA locally needs $k^3 \times (N/k) \approx O(k^2N)$ running time. To sum

up, the total running time needed for the batch mode algorithm is $O((k+1)N^2 + (k^2+k)N)$. If the proposed method is embedded in a manifold learning algorithm, then the running time complexity can be reduced to $O((k^2+k)N)$ in the case when the distance matrix and the neighborhood are already defined. This is a relatively small increase in the time complexity of a manifold learning algorithm which is always as high as $O(N^3)$.

For incremental learning, the neighborhood identification step needs $O(N/k)$ running time, whilst the local PCA consumes $O((k+1)^3)$ running time. Therefore, the total time complexity for incremental learning is $O((k+1)^3 + N/k)$.

IV. Experiments

The proposed algorithm was implemented with parameters $\alpha = 10$ and $\beta = 0.8$ for all the experiments.

In practice, it is found that noise contained in data is of low-dimension, except an additive white noise which is assumed to be in every component of the data vectors in $\mathbb{R}^D$. Thus, in practice, we only use variances of the first $\min(10, N-r+1)$ PCs in the noise part of data to estimate the variance of noise (see Eq. (3)).

Comparison is made among the k-$k/2$ NN method [18], the k-NNG method [26], the revised MLE (MLE in short) method [20], the C-PCA method and the L-PCA method, where the L-PCA method stands for the C-PCA method without the noise filtering procedure proposed in Subsection III-B. It should be noted that the results obtained by the MLE, k-$k/2$ NN and k-NNG methods are positive real numbers, while the L-PCA and C-PCA methods produce only integer ID estimations. In order to make a comparison among these results, we average the local ID estimations obtained with the C-PCA and L-PCA methods to provide a real ID estimation: $\hat{d} = \frac{1}{S}\sum_{i=1}^{S}\hat{d}_i$.

A. 10-Mobius data

The first data set is a 10-Mobius ring embedded in $\mathbb{R}^3$. Fig. 1(a) shows the scatter plot of the Mobius ring data set. As can be seen, the Mobius data points are lying on a highly nonlinear manifold with 1200 points uniformly distributing on the surface. Fig. 1(b) shows the results obtained by the five ID estimation algorithms against the neighborhood size ranging from 4 to 40. The MLE method is the most stable and accurate algorithm for all neighborhood sizes. All algorithms converge to the correct estimation. It seems that the L-PCA method does not diverge on this data set. This is possibly because the original dimensionality of data is low.
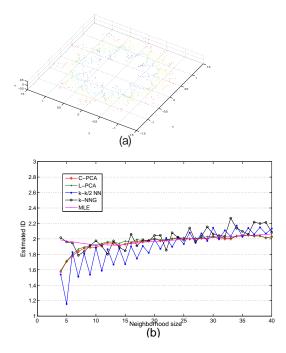

Fig. 1. (a) shows the scatter plot of the Mobius ring data set, and (b) shows the ID estimation results corresponding to the size of subregions.

B. Real world data sets

Our algorithm is compared with the MLE, k-$k/2$ NN and k-NNG methods on some benchmark real world data sets: the Isoface data set [4], the LLEface data set [5] and the MNIST '0' and '1' data sets [28].

The Isoface data set is comprised of 698 images of a head with the resolution 64×64. Some samples of the Isoface data set are shown in Fig. 2(a). In the experiments, each image is reshaped to a 4096-dimensional vector. It can be seen that the Isoface data set is under a three-dimensional movement: up-down, left-right and lighting changes. In [4], the Isomap algorithm estimated its ID as 3 using the projection approach. As can be seen from Fig. 2(b), corresponding to the neighborhood sizes from 4 to 40, the C-PCA estimator ranges from 2.3 to 3.5 and the MLE estimator ranges from 3.5 to 4.5. The estimation given by the k-NNG and k-$k/2$ NN methods is oscillating badly with the neighborhood sizes, so they are bot unstable. Since the L-PCA method can not filter out noise contained in data, it tends to overestimate the ID as the neighborhood size increases. This means that our noise filtering process plays a key role in the convergence of the C-PCA method.

The second data set is the LLEface data set, which contains 1965 samples in a 560-dimensional space (see Fig. 3(a) for some samples). From Fig. 3(b), it is seen that both the C-PCA and the MLE methods give a convergent ID estimation with the increase of the neigh-
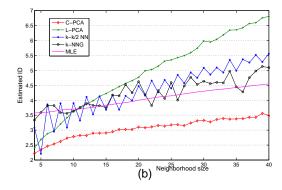
Fig. 2. (a) shows some samples of the Isoface data set. As can be seen, a head is under left-right, up-down and lighting changes. (b) presents the estimated ID of the Isoface data set.
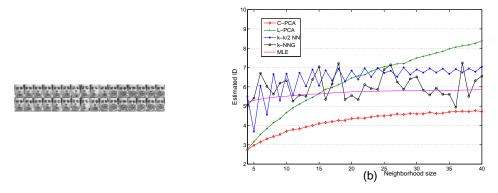

Fig. 3. (a) shows some samples of the LLEface data set and (b) plots the estimated ID of the LLEface data set against the neighborhood size.
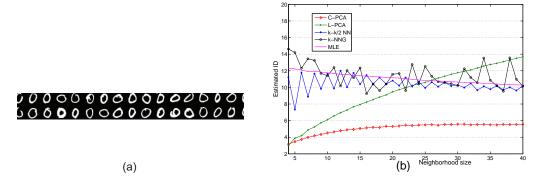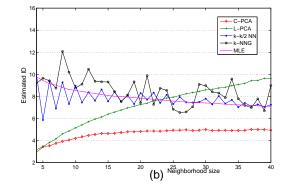

Fig. 4. (a) shows some samples of '0' in the MNIST data set and (b) gives the plot of the estimated ID of data '0' versus the neighborhood size.

borhood size, while the L-PCA, k-$k/2$-NN and k-NNG methods seem not convergent when the neighborhood size is increasing. The ID estimation given by the C-PCA method changes between 2.8 and 4.7 with the convergent estimation being 4.7, while the estimation result obtained by the MLE method changes gradually from 5.2 to 5.8 with a convergent estimation of 5.8.

We now consider two MNIST data sets: the set '0' and the set '1' (see Fig. 4(a) and Fig. 5(a) for some samples of these two data, respectively). The data set '0' contains

980 data points, while the data set '1' contains 1135 data points. It can be seen from Fig. 4(b) and Fig. 5(b) that all methods, except the L-PCA and k-NNG methods, converge with the increase of the neighborhood size. For the data set '0', it can be seen from Fig. 4(b) that the ID estimation given by the C-PCA method converges to 5.8 and the estimation given by both the MLE estimator and the k-$k/2$-NN estimator converges to 10. For the data set '1', Fig. 5(b) shows that the ID estimation obtained by

(a)

(b)

Fig. 5. (a) shows some samples of '1' in the MNIST data set and (b) presents the plot of the estimated ID of data '1' versus the neighborhood size.

the C-PCA method converges to 5.5 and the estimation provided with both the MLE method and the k-$k/2$-NN method converges to 7.2. Note that the result given by our method is in a big disagreement with the results given by other methods for the ID estimation of the data sets '0' and '1'. A digit '0' is usually represented as an ellipse which can be determined by the coordinates of its focus and its major and minor axes, so the ID of the data set '0' is likely to be 5. The number '1' can be considered as a line segment, which rotates from left to right, so a sensible ID estimation for the data set '1' may be between 4 and 5.

C. Noisy data sets

The traditional PCA algorithm is very sensitive to outliers, and the performance of PCA-based algorithms deteriorate rapidly if data points are sparse on a manifold such as the hand rotation data set [1]. As can be seen from Fig. 6(a), the hand is under a one-dimensional movement, so the data points can be considered as lying on a one-dimensional curve. The data set contains 481 image samples, and each sample is a vector in a 512480-dimensional space. Many outliers can be seen from its low-dimensional embedding by the Isomap algorithm (see Fig. 6(b)). Its ID estimation results with different methods are shown in Fig. 6(c).

Both the k-$k/2$ NN and k-NNG methods are sensitive to the choice of the neighborhood size and tend to overestimate the ID as the neighborhood size increases. On the other hand, the MLE estimator is more stable (see Fig. 6(c)). However, the minimum estimation of MLE method is 1.75, which is still higher than the ID of this data set. L-PCA method has the worst performance due to the outliers contained in the data set. The estimation

[1]CMU database: http://vasc.vi.cmu.edu/idb/html/motion/hand/index.html
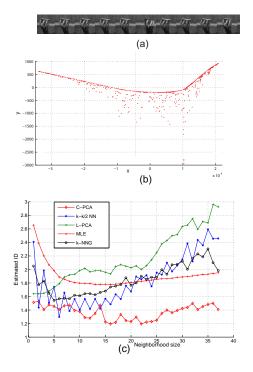

(a)

(b)

(c)

Fig. 6. (a) shows selected samples of the hand rotation data set, (b) shows the low dimensional embedding of hands rotation data sets by Isomap algorithm, (c) ID estimations of the hands rotation data set.

of the C-PCA method, which changes between 1.5 and 1.2, is the closest one to the correct ID of this data set.

We now transform the original 10-Mobius data in a 4-dimensional space using an Euclidean transformation. A random noise with mean 0 and variance 0.2 is added to the transformed data. The ID estimation results with different algorithms are given in Fig. 7. As can be seen from Fig. 7, the ID estimation given by the C-PCA method is the closest one to the correct ID of this noised 10-Mobius data set. The other algorithms tend to overestimate the ID of the noised data set. The estimation obtained by the L-PCA method is a little higher than that given by the C-PCA due to the effect of noise.
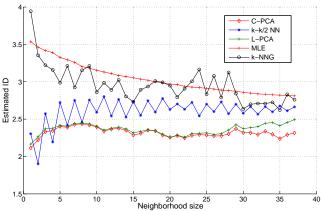
Fig. 7. ID estimations of the noised Mobius data set.

V. Conclusion

In this paper, we proposed a new ID estimation method based on PCA. The proposed algorithm is simple to implement and gives a convergent ID estimation corresponding to a wide range of neighborhood sizes. It is also convenient for incremental learning. Experiments have shown that the new algorithm has a robust performance.

Acknowledgment

References

[1] T.M. Buzuga, J.V. Stammb, G. Pfister, Characterising experimental time series using local intrinsic dimension, Physics Letters A202 (1995), 183-190.

[2] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford Univ. Press, Oxford, 1995.

[3] J. Friedman, T. Hastie, R. Tibshirani, The Elements of Statistical Learning - Data Mining, Inference and Prediction, Springer, Berlin, 2001.

[4] J.B. Tenenbaum, V. de Sliva, J.C. Landford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000), 2319-2323.

[5] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000), 2323-2326.

[6] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and representation, Neural Computation 15 (2003), 1373-1396.

[7] H.S. Seung, D.D. Lee, The manifold ways of perception, Science 290 (2000), 2268-2269.

[8] I.T. Jolliffe, Principal Component Analysis, Springer, Berlin, 1989.

[9] K. Fukunaga, D.R. Olsen, An algorithm for finding intrinsic dimensionality of data, IEEE Transactions on Computers 20 (1971), 176-183.

[10] J. Bruske, G. Sommer, Intrinsic dimension estimation with optimally topology preserving maps, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998), 572-575.

[11] T. Hastie, W. Stuetzle, Principal curves, Journal of the American Statistical Association 84 (1988), 502-516.

[12] S. Chatterjee, M. R. Yilmag, Chaos, fractals and statistics, Statistical Science 7 (1992), 49-68.

[13] P. Grassberger, I. Procaccia, Measuring the strangeness of strange attractors, Physica D9 (1983), 189-208.

[14] B. Kegl, Intrinsic dimension estimation using packing numbers, Advances in Neural Information Processing Systems 16 (2002), 681-688.

[15] T. Lin, H. Zha, Riemannian manifold learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008), 796-809.

[16] P.J. Verveer, R.P.W. Duin, An evaluation of intrinsic dimensionality estimators, IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (1995), 81-86.

[17] M. Fan, H. Qiao, B. Zhang, Intrinsic dimension estimation of manifolds by incising balls, Pattern Recognition 42 (2009), 780-787.

[18] A.M. Farahmand, C. Szepesvari, J.Y. Audibert, Manifold-adaptive dimension estimation, in: Proceedings of the 24th Annual International Conference on Machine Learning, 2007, pp. 265-272.

[19] E. Levina, P.J. Bickel, Maximum likelihood estimation of intrinsic dimension, Advances in Neural Information Processing Systems 18 (2004), 777-784.

[20] D.J.C. MacKay, Z. Ghahramani, Comments on 'Maximum likelihood estimation of intrinsic dimension' by E. Levina and P. Bickel, see
http://www.inference.phy.cam.ac.uk/mackay/dimension/, 2005.

[21] M. Raginsky, S. Lazebnik, Estimation of intrinsic dimensionality using high-rate vector quantization, Advances in Neural Information Processing Systems 19 (2005), 352-356.

[22] F. Camastra, Data dimensionality estimation methods: a survey, Pattern Recognition 36 (2003), 2945-2954.

[23] M. Hein, J.Y. Audibert, Intrinsic dimensionality estimation of submanifolds in $\mathbb{R}^d$, in: Proceedings of the 22nd International Conference on Machine Learning (ed. Morgan Kaufmann), 2005, pp. 289-296.

[24] S.W. Cheng, Y.J. Wang, Z.Z. Wu, Provable dimension detection using principal component analysis, in: Proceedings of the 21th Annual Symposium on Computational Geometry, 2005, pp. 208-217.

[25] S.W. Cheng, M.K. Chiu, Dimension detection via slivers, in: Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms, 2009, pp. 1001-1010.

[26] J.A. Costa, A.O. Hero, Geodesic entropic graphs for dimension and entropy estimation in manifold learning, IEEE Transactions on Signal Processing 52 (2004), 2210-2221.

[27] J.A. Costa, A.O. Hero, Estimating local intrinsic dimension with k-nearest neighbor graphs, IEEE Transactions on Statistical Signal Processing 30 (23) (2005), 1432-1436.

[28] Y. Le Cun, L. Bottou, Y. Bengio, H. Patrick, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (1998), 2278-2324.

[29] K.W. Pettis, T.A. Bailey, A.K. Jain, R.C. Dubes, An intrinsic dimensionality estimator from near-neighbor information, IEEE Transactions on Pattern Analysis and Machine Intelligence 1 (1979), 25-37.