Research on the Construction Algorithm of Principal Curves

Lianwei Zhao[1], Yanchang Zhao[2], Siwei Luo[1], and Chao Shao[1]

[1] School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
lw_zhao@126.com
[2] Faculty of Information Technology, University of Technology, Sydney, Australia
yczhao@it.uts.edu.au

Abstract. Principal curves have been defined as self-consistent, smooth, one-dimensional curves which pass through the middle of a multidimensional data set. They are nonlinear generalization of the first Principal Component. In this paper, we take a new approach by defining principal curves as continuous curves based on the local tangent space in the sense of limit. It is proved that this new principal curves not only satisfy the self-consistency property, but also are the unique existence for any given open covering. According to the new definition, a new practical algorithm for constructing principal curves is given too. And the convergence properties of this algorithm are analyzed. The new construction algorithm of principal curves is illustrated on some simulated data sets.

Areas. data mining, machine learning

Research on the Construction Algorithm of Principal Curves

Lianwei Zhao[1], Yanchang Zhao[2], Siwei Luo[1], and Chao Shao[1]

[1] School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
lw_zhao@126.com
[2] Faculty of Information Technology, University of Technology, Sydney, Australia
yczhao@it.uts.edu.au

Abstract. Principal curves have been defined as self-consistent, smooth, one-dimensional curves which pass through the middle of a multidimensional data set. They are nonlinear generalization of the first Principal Component. In this paper, we take a new approach by defining principal curves as continuous curves based on the local tangent space in the sense of limit. It is proved that this new principal curves not only satisfy the self-consistent property, but also are the unique existence for any given open covering. Based on the new definition, a new practical algorithm for constructing principal curves is given. And the convergence properties of this algorithm are analyzed. The new construction algorithm of principal curves is illustrated on some simulated data sets.

1 Introduction

Finding low-dimensional manifold embedded in the high-dimensional space is a fundamental problem in the field of data mining, pattern recognition and computer vision. The research on this problem started as linear, and then as nonlinear parametric model, at last generated to the nonlinear non-parametric model. Meantime, there bring many mathematic problems, such as theory foundation, approximation algorithm and so on. Principal Curves are the nonlinear generalization of first principal components, and have been defined as smooth one-dimensional curves, which pass through the middle of a multidimensional data set. Although the description is intuitional, there are different definitions about the middle of data distribution. Principal curves were firstly introduced by Hastie and Stuetzle [1], and have been defined as satisfying the self-consistency property (HSPC). Tharpey and Flury[2] generalize the concept of self-consistency to random vectors, and then construct a unified theoretical basis for principal components, principal curves and surfaces, principal points, principal variables, and other statistical methods, and show the relationships between the various method. Tibshirani[3] give an alternative definition of principal curves based on a mixture model, then carry out the estimation through EM algorithm. This model, however, has inducted parameters, and can not give the method to remove the bias of estimation, and in practice does not satisfy the

property of self-consistency. Kégl et al. [4] provide a new definition for principal curves with bounded length, and show that such curves exist for any distribution with bounded second moment. They also give an algorithm to implement the proposals, and calculated rates of convergence of the estimators. Due to the length constraint, the treatment does not encompass the case of classical principal component analysis. Delicado[5] introduces a new definition of principal curves based on the principal oriented points, prove the existence of principal curves passing through these points, and then propose an algorithm to find the principal curves, but all the arguments are based on conditional expectation and variance. Chang[6] propose an unified model as probabilistic principal surface (PPS) to address a number of issues associated with current algorithms using a manifold oriented covariance noise model, based on the generative topographical mapping (GTM), which can be viewed as a parametric formulation of SOM. In the past twenty years, the research on the principal curves around the middle property of data distribution has got excited progress, however, there have some open problems. For example, the existence of principal curves cannot be guaranteed for any distributions, and theoretical analysis is not as straightforward as with parametric models due to its nonparametric formulation. Duchamp and Stuetzle[7,8], José L. Martínez-Morales[9] study the differential geometry property of principal curves in the plane, find that the largest and the smallest principal components are extrema of the distance of expected distance from the data points, but all the principal curves are saddle points. This means that cross-validation can not be used to choose the complexity of principal curves estimates. By solving differential equation, they find that there are oscillating solutions and principal curves will not be unique. These conclusions indicate that the middle of data distribution lacks the sufficient theoretic support.

Therefore, we have to turn back to consider the problem of nonlinear generalizations. There are two different ways in technology: one is that we can assume the data set obey some kind of distribution, and then find the statistical distribution model which is the best of the intrinsic structures describing this distribution. Manifold fitting, principal curves and GTM are the representative algorithms. Another is to transform the input data space, and then computer the linear component, such as kernel PCA. Principal component analysis is a widely used tool in multivariate data analysis for purposes such as dimension reduction and feature extraction. Now that principal curves are the nonlinear generalization of principal components, they can be used for reference more idea from linear PCA. PCA can be used to project the high-dimensional observed data to low-dimensional principal subspace, and the preconditions is that data set can be embedded in the global linear or approximation linear low-dimensional sub-manifold. So if the sub-manifold is nonlinear, PCA can not preserve the local information. Eliminating the statistical redundancy among the components of high-dimensional data with little information loss, is the main goal of finding low-dimensional representation. So in this problem, although the data distribution is nonlinear in global, can we think it as linear in local? And is it feasible in theoretic and algorithm?

In the following of this paper, we firstly introduce definition and construction algorithm of HSPC and discuss the property of self-consistency in Section 2. Then in Section 3, according to the relation between the local tangent space and principal components, a new definition of principal curves is given in the sense of limit. We

also prove that this principal curves satisfy the self-consistency property, and the existence of that curves for any given open covering. Based on this definition, a constructing algorithm of principal curves is proposed in Section 4. Our experimental results on simulate data sets are given in Section 5. Conclusions are provided in the last section.

2 Definition of HSPC and Self-consistency Property

2.1 Definition and Construction Algorithm of HSPC

Problem Description: Consider a multivariate random variable $X=(X_1, X_2, \cdots, X_p)$ in R^p with density function $p_X(x)$ and a random sample from X, named $x_1, x_2, \cdots, x_n$, then how to find a one-dimensional smooth curves $f(\lambda)$, which pass through the middle of X ?

The first principal component can be viewed as the straight line which best fits the clouds of data. Principal curves were firstly introduced by Hastie to formalize the notion of curves passing through the middle of a dataset.

Definition (HSPC): let $f(\lambda)$ be smooth curves in R^p, parametrized by $\lambda \in R$, and for any $\lambda \in R$, let projection index $\lambda_f(x)$ denote the largest parameter value λ for which the distance between x and $f(\lambda)$ is minimized, i.e., $\lambda_f(x)=\sup_{\lambda}\left\{\lambda : \|x-f(\lambda)\| = \inf_{\mu}\|x-f(\mu)\|\right\}$. Then principal curves are the curves satisfying the self-consistency property $E(X \mid \lambda_f(X)=\lambda)=f(\lambda)$.

Hastie has proved that the project index should be a random variable, and found the principal curves have the same property as principal component. But according to this definition, HSPC cannot be the self-intersecting curves. Given the density function of X, HS principal algorithm for constructing principal curves is given in the following:

Step 1: Set $f^{(0)}(\lambda)$ be the first principal component line for X, and set $j=1$;

Step 2: Define $f^{(j)}(\lambda)=E(X \mid \lambda_{f^{(j)}}(X)=\lambda)$;

Step 3: Compute $\lambda_{f^{(j)}}(x)=\max\left\{\lambda : \|x-f^{(j)}(\lambda)\| = \min_{\mu}\|x-f^{(j)}(\mu)\|\right\}$, for all $x \in R^d$;

Step 4: Compute $\Delta(f^{(j)})=E\left\|(X-f^{(j)}(\lambda_{f^{(j)}}(X))\right\|$. If $\left|\Delta(f^{(j)})-\Delta(f^{(j-1)})\right| < threshold$, then stop. Otherwise, let $j=j+1$ and go to Step 2.

In practice, the distribution of X is often unknown, but the data set consisting of n samples from X is known, so the expectation in step 2 can be substituted by a smoother or non-parametric regression estimation.

2.2 Self-consistency Property

In HSPC, self-consistency is introduced to describe the property that each point on the smooth curves is the mean of all points projected onto it. Self-consistency is the fundamental property of principal curves, and then is generalized to define the self-consistent random vectors.

Definition 2: A random vector Y is self-consistent for X if each point in the support of Y is the conditional mean of X, namely $E(X|Y)=Y$.

For jointly distributed random vectors Y and X, the conditional mean $E(Y|X)$ is called the regression of Y on X. For the regression equation $Y=f(X)+\varepsilon$, where X is m-dimensional random variable, $f(\cdot)$ a function from R^m to R^n, ε is a n-dimensional random vector independent of X ,and $E(\varepsilon)=0$. Then $E(Y|f(X))=f(X)$, that is, $f(X)$ is self-consistent for Y.

In practice, because we only have the finite data set, so the point projected on the principal curves is only one at most. Therefore Hastie introduces the conception of neighborhood and defines the point as conditional expectations of data set projected in the neighborhood. This definition also agrees with the mental image of a summary.

3 Definition of Principal Curves Based on Local Tangent Space

Consider $f(\lambda)$ is continuously differentiable curve, where $\lambda \in \mathrm{supp}(\lambda)$. A Taylor series approximation about λ_0 is $f(\lambda)=f(\lambda_0)+J_f(\lambda_0)(\lambda-\lambda_0)+\mathrm{O}(\|\lambda-\lambda_0\|^2)$, where $J_f(\lambda_0)$ is the Jacobian matrix. The tangent space of $f(\lambda)$ on λ_0 will be spanned by the column vectors of $J_f(\lambda_0)$. So the points in the neighborhood of $f(\lambda_0)$ can be approximated by $f(\lambda)\approx f(\lambda_0)+J_f(\lambda_0)(\lambda-\lambda_0)$. A new definition of principal curves is presented in the following:

Definition 3: let $f(\lambda)$ be a smooth curves in R^p, parametrized by $\lambda \in R$. For any $\lambda \in R$, a cluster of neighborhood $\{B(\lambda_1,\delta_1),B(\lambda_2,\delta_2),\cdots,B(\lambda_k,\delta_k)\}$ is an open covering of λ, where $\lambda_i \in \lambda, i=1,2,\cdots,k$, ξ_i is the local tangent space vector of $f(\lambda)$ in the $B(\lambda_i,\delta_i)$. Let $P_{\xi_i}(x), x\in B(\lambda_i,\delta_i)$ be the projection on the ξ_i. If $\lim_{\delta\to 0}\sum_{i=1}^{k} P_{\xi_i}(x)\delta_i$ is existent, then $f(\lambda)=\lim_{\delta\to 0}\sum_{i=1}^{k} P_{\xi_i}(x)\delta_i$ is the principal curves, and satisfies the self-consistency property, where $\delta=\max_{i=1,2,\cdots,k}\{\delta_i\}$.

Some remarks on the above definition is given in the following.

a). Definition of principal curves is in the sense of limit.

Supposed random samples $x_1,x_2,\cdots,x_n$ is iid, and embedded on the smooth manifold M in R^p, and $x=f(\lambda)$, $d\in R^m$,where, $m<<p$. The mean of samples is $\bar{x}$, the

covariance matrix is $Cov(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$. Wanli Min et al.[11] have proved that the eigenvectors of $Cov(X)$ can construct the tangent space of manifold M on $\bar{x}$. This indicates that the local tangent space can be approximated by the eigenvectors of the covariance matrix of samples, and in each local neighborhood, the topology structure can be preserved.

So for any $\lambda_i \in \lambda, i = 1,2,\cdots,k$, its tangent vector can be approximated by the first eigenvectors ξ_i of covariance matrix of data points in the $B(\lambda_i, \delta_i), i = 1,2,\cdots,k$.

So we have $\sum_{i=1}^{k} P_{\xi_i}(x)\delta_i \approx \sum_{i=1}^{k} J_f(\lambda_i)\delta_i$

Because $J_f(\lambda_i) = \frac{f(\lambda) - f(\lambda_i)}{\lambda - \lambda_i} - \frac{O(\|\lambda - \lambda_i\|^2)}{\lambda - \lambda_i}$, $\lim_{\lambda \to \lambda_0} J_f(\lambda_i) = \lim_{\lambda \to \lambda_0} P_{\xi_i}(x)$ is existent.

Then $\lim_{\delta \to 0}\sum_{i=1}^{k} P_{\xi_i}(x)\delta_i = \lim_{\delta \to 0}\sum_{i=1}^{k} J_f(\lambda_i)\delta_i = \int_{\lambda} J_f(\lambda)d\lambda = f(\lambda)$, where $\delta = \max_{i=1,2,\cdots,k}\{\delta_i\}$.

b). Satisfying the self-consistency property

Self-consistency is the fundamental property of principal curves, in the following we will show how this definition satisfying this property.

Suppose for given any neighborhood $B(\lambda_i, \delta_i)$, P is the orthogonal projection from R^p to its q linear subspace M . If $Y_i = (I - P)E(X_i) + PX_i$ is self-consistency for X_i , then M is spanned by the q eigenvectors of $Cov(X_i)$. If $P = \xi\xi^{'}$, where ξ is the first principal components of $Cov(X_i)$, the $Y_i = (I - \xi\xi^{'})E(X_i) + \xi\xi^{'}X_i$ is self-consistency for X_i .

For any point $f(\lambda^*)$ on the $f(\lambda)$, $\{B(\lambda_1, \delta_1), B(\lambda_2, \delta_2), \cdots, B(\lambda_k, \delta_k)\}$ is an open covering for λ , so there exist a $\lambda_i \in \lambda, i = 1,2,\cdots,k$, $f(\lambda^*) \in Y_i \subset B(\lambda_i, \delta_i)$, satisfying $E(X \mid \lambda_f(X) = \lambda^*) = f(\lambda^*)$. Therefore, $f(\lambda)$ is self-consistency.

c). For any given covering, there exists a principal curves which minimizes LRE

Suppose random vector X and Y , for any function g , $E\|X - E(X \mid Y)\| \le E\|X - g(Y)\|$. Taking g to be the identity, then $E\|X - E(X \mid Y)\| \le E\|X - Y\|$. Thus if $Y = E(X \mid Y)$, Y is local optimal for approximating X .

For the observed data sets, in every neighborhood $B(\lambda_i, \delta_i), i = 1,2,\cdots,k$, the local reconstruction error LRE ($LRE_{ki} = \sum_{x_{ij} \in B(\lambda_i, \delta_i)} \|x_{ij} - \hat{x}_{ij}\|_2^2$) is minimal if and only if $\hat{x}_{ij} = \bar{x} + \xi_h \xi_h^T (x_{ij} - \bar{x})$, where ξ_k is the eigenvectors of the input covariance matrix corresponding to its largest eigenvalue.

Hence, for any given open covering, there exist curves which minimize the local reconstruction error.

4 Construction Algorithm of Principal Curves

In this section, we give a construction algorithm of principal curves based on the local tangent space, according to Definition 3. The algorithm of construct principal curves is described as following.

Step 1: Let neighborhood $\{B_k(\lambda_1,\delta_1), B_k(\lambda_2,\delta_2),\cdots, B_k(\lambda_k,\delta_k)\}$ be an open covering for sample;

Step 2: For the sample points $x_{i1}, x_{i2},\cdots, x_{in}$ in $B_k(\lambda_i,\delta_i), i=1,2,\cdots k$, compute the $E_{ki}(x)$, $Cov_{ki}(x)$, and $V_{kij}=(v_{i1}, v_{i2},\cdots, v_{ij})$;

Step 3: Let $V_{ki}=\min\{V_{(k-1)l}{}^T v_{im}\}, m=1,\cdots, j$, where $V_{(k-1)l}$ denotes the principal eigenvector of covariance matrix of sample points in $B_{k-1}(\lambda_l,\delta_l)$, and $B_k(\lambda_i,\delta_i)\subset B_{k-1}(\lambda_l,\delta_l)$. Compute the projection of data point x_{ki} on V_{ki}, and reconstruction $\hat{x}_{ki}$;

Step 4: Connect $\hat{x}_{ki}$, and use the method for local smooth interpolation;

Step 5: Compute global reconstruction error GRE_k. If $GRE_k - GRE_{k-1} < threshold$, then stop. Otherwise, let $k=k+1$, and go to Step 1.

Note about the convergence properties of this algorithm in the following:

For every $LRE_{ki}\geq 0$, the global least reconstruction error is

$$GRE_k=\sum_{i=1}^{k} LRE_{ki}=\sum_{i=1}^{k}\sum_{x_{ij}\in B(\lambda_i,\delta_i)}\left\|x_{ij}-\hat{x}_{ij}\right\|_2^{\,2}=\sum_{i,j}\left\|x_{ij}-\bar{x}+\xi_h\xi_h{}^T(x_{ij}-\bar{x})\right\|_2^{\,2}$$, then

$$\begin{aligned} GRE_k-GRE_{k-1} &= \sum_{i,j}\left\|x_{ij}-\bar{x}+\xi_h\xi_h{}^T(x_{ij}-\bar{x})\right\|_2^{\,2}-\sum_{i,j}\left\|x_{ij}-\bar{x}+\xi_{h-1}\xi_{h-1}{}^T(x_{ij}-\bar{x})\right\|_2^{\,2} \\ &\leq \sum_{i,j}\left(\left\|x_{ij}-\bar{x}+\xi_h\xi_h{}^T(x_{ij}-\bar{x})-x_{ij}+\bar{x}-\xi_{h-1}\xi_{h-1}{}^T(x_{ij}-\bar{x})\right\|_2^{\,2}\right) \\ &= \sum_{i,j}\left(\left\|(\xi_h\xi_h{}^T-\xi_{h-1}\xi_{h-1}{}^T)(x_{ij}-\bar{x})\right\|_2^{\,2}\right) \end{aligned}$$

For any continuous differential function $f(\lambda)$ and $\lambda_0\in \mathrm{supp}(\lambda)$, where λ is a compact subset of R., $B(\lambda_0,\delta_k)$ and $B(\lambda_0,\delta_{k-1})$ are two neighborhood of λ_0, and $B(\lambda_0,\delta_{k-1})\subset B(\lambda_0,\delta_k)$. For any $\lambda_1\in B(\lambda_0,\delta_k)$ and $\lambda_2\in B(\lambda_0,\delta_{k-1})$,

$$\lim_{\lambda_1\to\lambda_0}\left|\frac{f(\lambda_1)-f(\lambda_0)}{\lambda_1-\lambda_0}-\frac{f(\lambda_2)-f(\lambda_0)}{\lambda_2-\lambda_0}\right|=\left|\xi_k-\xi_{k-1}\right|=0$$

Therefore, $GRE_k-GRE_{k-1}\to 0$.

5 Experimental Results

To test the algorithm presented above, we conducted experiments on several artificial data sets. Consider a random sample $x_1, x_2,\cdots, x_n$ from a multi-dimensional random

variable X , suppose that a nonlinear curves is a good summary of the structure of the distribution of X and we try to recovering the curves from the observed samples $x_1, x_2, \cdots, x_n$.

5.1 Experiments on Continuous Function without Noise

Consider $y = \sin(x), x \in [-2,2]$, the number of selected points is 400, and the number of neighborhood is 1, 2, 3, 10, respectively. In Fig. 1, we illustrate several stages of the principal curve constructing. And the result is promising, when k=10, we can see that the principal curves constructed with the proposed algorithm have approximated to the origin continuous functions.

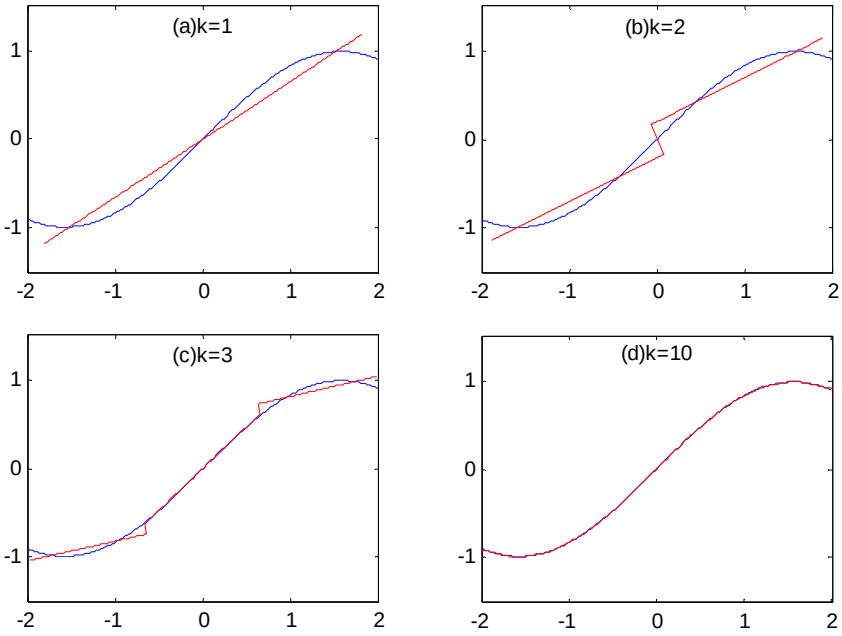

Fig. 1. Principal Curves. The data was generated by continuous function (sinusoid): (a)k=1, (b)k=2, (c)k=3, (d)k=10

5.2 Experiments on Gaussian Distributions

Consider the two independent Gaussian distribution and randomly selected 100 points. We get the principal curves with k=2 and 5, as Fig. 2 shows. From these results, we can see the principal curves approximate to the principal component.

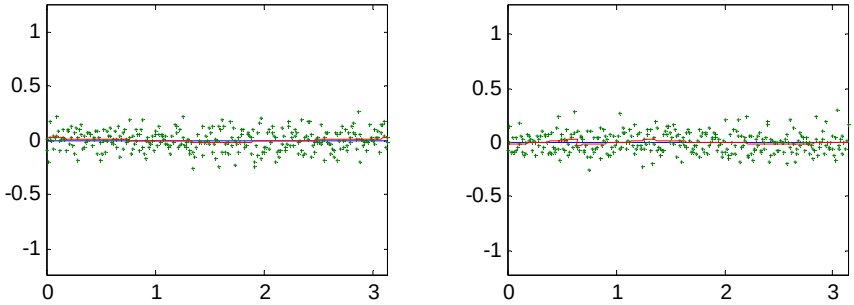

Fig. 2. The Principal Curves from elliptical distribution, with k=2(left) and k=5(right) respectively

5.3 Experiments on Continuous Function with Noise

Consider the $x = \sin\theta, y = \cos\theta, \theta \in [0, \pi]$, randomly select 200 points, and add independent Gaussian noise $\varepsilon_i \sim N(0,0.1)$. Let k=1,2,4,40 respectively and we illustrate several stages of the principal curve constructing in fig. 3. The result is also promising, when k=40, we can see that the principal curves constructed with the proposed algorithm have approximated to the origin continuous functions with noise.

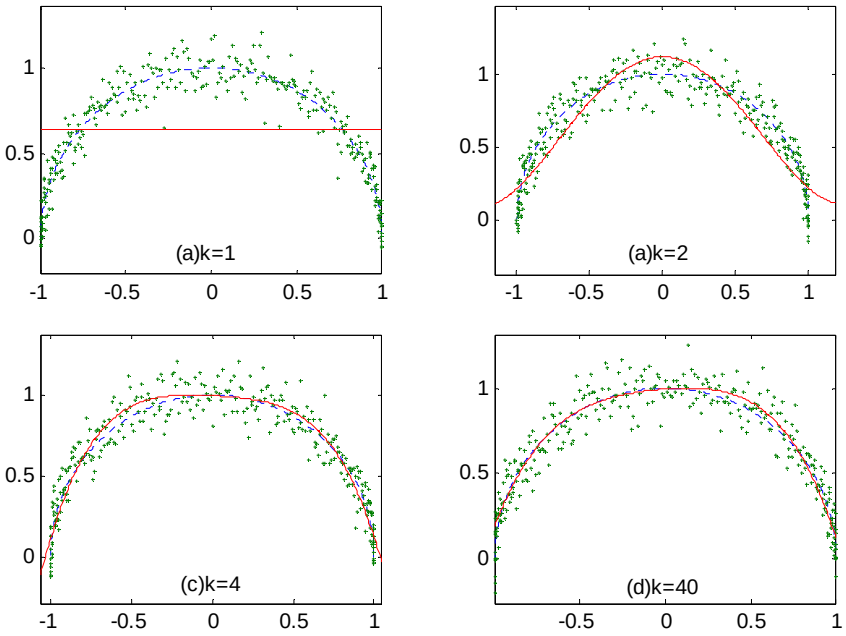

Fig. 3. The Principal Curves. The data was generated by adding independent Guassian noise on a half circle (a)k=1, (b)k=2, (c)k=4, (d)k=40

And in the fig.4, we give other two experiments results. Due to limited space we cannot present exhaustive experimental results but just some illustrations here.

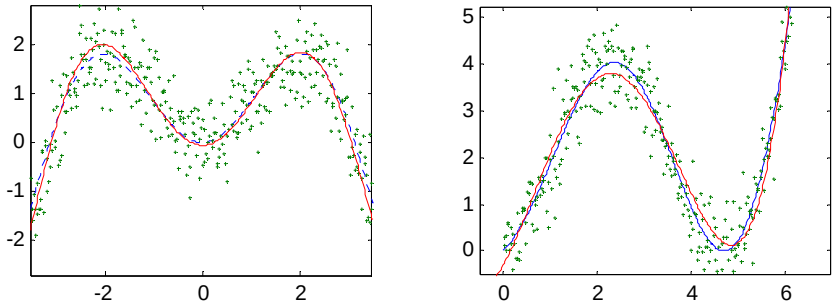

Fig. 4. The Principal Curves. The data was generated by adding independent Guassian noise on continuous function

6 Conclusions

For high-dimensional random vector, it is very important to find an approximation whose support is a low-dimensional manifold. Principal curves can be regard as one dimension principal manifold, and have been used for dimension reduction and pattern classification. In this paper based on the local tangent space, we give a new definition of principal curves in the sense of limit, and prove it is self-consistency. We also show for any given open covering, this principal curves exists. According to the definition, we give a construction algorithm of principal curves. Experimental results show that we can approximate to the true principal curves. In this paper we suppose the domain of principal curves is compact set and can be covered by finite open covering. How to find an open covering for more complex data set will be future work.

Acknowledgements

The research is supported by national natural science foundations of china (60373029).

References:

1. Hastie T and Stuetzle W. Principal Curves. Journal of the American Statistical Association. 1989,84: 502-516.
2. Tarpey T and Flury B. Self-consistency: A fundamental concept in statistics. Statistical Science. 1996,11 (3): 229-243.
3. Tibshirani R. Principal Curves Revisited. Statistics and Computation. 1992,2:183-190.

4. Kégl B, Krzyzak A, Linder T and Zeger K. Learning and design of principal curves. IEEE Trans. on Pattern Analysis and Machine Intelligence. 2000,22 (3): 281-297.
5. Delicado P. Another Look at Principal Curves and Surfaces. Journal of Multivariate Analysis.2001,77:84-116.
6. Chang Kui-yu and Joydeep Ghosh. A Unified Model for Probabilistic Principal Surfaces. IEEE Trans. on Pattern Analysis and Machine Intelligence. 2001,23(1):22-41.
7. Duchamp T and Stuezle W. Geometric Properties of Principal Curves in the Plane. Robust Statistics. Data Analysis, and Computer Intensive Methods. 1995:135-152.
8. Duchamp T and Stuetzle W. Extremal Properties of Principal Curves in the Plane. Annals of Statistics. 1996,24 (4): 1511-1520.
9. José L. Martínez-Morales. Extremal Properties of Principal Embeddings. J. Math. Pures Appl.1999,78: 913-923.
10. ZHANG Jun-ping and WANG Jue. An Overview of Principal Curves. Chinese Journal of Computers. 2003,26(2):129-146.
11. Min Wanli, Lu Ke, He Xiaofei. Locality Pursuit Embedding. Pattern Recognition, 2004,37: 781-788.
12. Jos Koetsier, Ying Han, Colin Fyfe. Twinned principal curves, Neural Networks. 2004,17: 399-409.
13. Kambhatla N and Leen T.K. Dimension reduction by local principal component analysis. Neural Computation.1997, 9: 1493-1516.