Dimension Reduction of Microarray Data Based on Local Principal Component

Ali Anaissi[#1], Paul J. Kennedy[#2],Madhu Goyal[#3]

Abstract—Analysis and visualization of microarraydata is veryassistantfor biologists and clinicians in the field of diagnosis and treatment of patients. It allows Clinicians to better understand the structure of microarray and facilitates understanding gene expression in cells. However, microarray dataset is a complex data set and has thousands of features and a very small number of observations. This very high dimensional data set often contains some noise, non-useful information and a small number of relevant features for disease or genotype. This paper proposes a non-linear dimensionality reduction algorithm Local Principal Component (LPC) which aims to maps high dimensional data to a lower dimensional space. The reduced data represents the most important variables underlying the original data. Experimental results and comparisons are presented to show the quality of the proposed algorithm. Moreover, experiments also show how this algorithm reduces high dimensional data whilst preserving the neighbourhoods of the points in the low dimensional space as in the high dimensional space.

Keywords—Linear Dimension Reduction;Non-Linear Dimension Reduction; Principal Component Analysis; Biologists.

I. Introduction

DIMENSIONALITY reduction is one of the most effective and essential tools in the microarray domain. It aims to reduce, understand and visualize the structure of complex data sets by transforming a high-dimensional data set into a lower dimensional data set which represents the most important variables that underlie the original data. This significant and important tool attracts many researchers working in the field of bioinformatics and deals with gene expression data sets to work on the dimensionality reduction [1], [2], [3].

High dimensionality with low numbers of observations is one of characteristics of gene expression data sets. One reason for this is because microarray experiments are expensive to produce many replications. As a result, analysis and visualization is difficult in practice and becomes an obstacle for clinicians and biologists in the field of diagnosis and treatment of patients such as childhood leukaemia sufferers [4]. Visualizing high dimensional data and extracting the effective dimension of the data set are two important outcomes achieved by dimensionality reduction.

Biologists and clinicians may be able to better understand the structure of a complex microarray data set and the gene expression in cells when reduced and visualized in 3D or 2D. Moreover, dimensionality reduction is an essential tool in the microarray domain in order to extract the effective dimension of the data set and reduce high dimensional data into more easily handled low dimensional data [5]. For example, due to the curse of dimensionality you could not directly find and retrieve similar data points for a given data point in a very high dimensional space without applying a dimensionality reduction technique as a pre-processing step for retrieving process.

Several algorithms and techniques have been proposed for dimensionality reduction. Principal component analysis (PCA) [15] is one of the most popular and widely used techniques. PCA is considered as a linear method and very simple effective tool but it is not efficient for high dimensional and complex data set. This is due to the fact that PCA can not retrieve precisely the true latent variables of complex and non-linear data sets [6]. Data in a very high dimensional space often exists in a lower dimensional nonlinear manifold. With this kind of data, the intrinsic nonlinear structure could not be found through a linear dimension reduction technique. Another drawback of PCA is that the size of the covariance matrix is proportional to the dimensionality of the data-points. In microarray datasets, where the number of variables is muchgreater than the number of data points (a typical microarray dataset would have a 150 data points with thousands of variables), the computation of eigenvalues and eigenvectors is costly and might be impracticable for the covariance matrix.

In order to overcome the drawback of linear dimensionality reduction in a very high dimensional dataset, several non-linear dimensionality reduction methods have been developed. Non-Linear Dimensionality Reduction methods are often more powerful than linear ones, because the connection between the latent variables and observed ones may be much richer than simple matrix multiplication [6].

A recent development of non-linear algorithm is Local Linear Embedding (LLE) [2]. LLE is efficient and powerful for dimensionality reduction among the other algorithms [6], [7], [8]. However, this algorithm has a good performance when applied on protein structure description.

Local Tangent Space Analysis (LTSA) is another nonlinear dimensionality reduction technique that describes local properties of the high-dimensional data using the local tangent space of each data point [10]. This technique has been successfully applied on microarray data. However, it requires

Center of Quantum Computation and Intelligent Systems (QCIS), Faculty of Engineering and Information Technology (FEIT),University of Technology, Sydney (UTS) .P.O. Box 123, Broadway, NSW 2007.Australia
1-e-mail: aanaissi@eng.uts.edu.au,
2- e-mail: Paul.Kennedy@uts.edu.au
3- e-mail: madhu@it.uts.edu.au

$k \geq d$ where k is the number of nearest neighbourhoods and d is the dimensionality target. As a result, LTSA is good in visualization.

In this paper, a nonlinear dimension reduction algorithm is proposed to handle the curse of dimensionality of microarray data. Local Principal Component (LPC) is a new algorithm for nonlinear dimension reduction. The algorithm is based on the first principal component of the local neighbourhood of each data point. The idea behind this algorithm is that the drawback of PCA is when it applied on a non linear and folded data. However, if we apply PCA on a local neighbourhood of each data point, these local data points might not folded and has a linear structure. For example, in a Swiss roll data, 100 data points at least are required to have folded shape with the non linear structure. The experiments show LPC outperforms PCA in reducing the dimension of non-linear structures and visualization performance.

The rest of this paper is organized as follows. Section II introduces the algorithm Local Principal Component.Section III introduces the different datasets that used in this study.Section IV discusses the quality of the algorithm and error estimation.We will discuss the validation of this algorithm in Section V by applying LPC on Iris, Swiss roll and microarray dataset.SectionVI presents a comparative review with other similaralgorithms. In Section VII, we draw conclusions about the results and present some of the future work.

II. Algorithm of LPC

This algorithm takes as input $\mathbf{X}\in \boldsymbol{R}^{m\times N}$and produces output$\mathbf{Y}\in\boldsymbol{R}^{d\times N}$where $d<m$ is the dimensionality of the embedding input vector **X** in the low dimensional space **Y**. Four steps are involved in this algorithm. The first step is to compute the neighbors for each data point. For that, we determine the k-nearest neighbors for each data point based on the Euclidean distance. As with many non NLDR algorithms, the quality of dimensionality reduction is sensitive to the value of parameter k which should be carefully chosen. Otherwise the result will be exposed to the loss of quality. If this parameter is tuned with a very high value, the algorithm will loose its nonlinear character and act as a linear dimensional reduction. On the other hand, if the value is too small, the data points will be above each other and the mapping will not reflect any global properties [11].After computing the k nearest neighbors for each data point, N cells or matrices are created with $d*k$ size.

The second step is to determine the first principal component for each matrix by solving the eigenvalues and eigenvectors problem of the covariance matrix.

The third step is to calculate the orthogonal projection of the first eigenvector before storing them in a square matrix M $\in \mathrm{R}^{N*N}$based on the indexes of the neighbourhoodsindices obtained from the first step..

The final step is to calculate the embedding coordinates Y using the **M** matrix and find the spectral embedding vector using the eigenvectors of this matrix. This task is achieved by solving the global eigenvalues and eigenvectors of the squared matrix **M**.

These steps are accomplished using the following algorithm:

Step 1. For each i = 1, · · ·, N find the k local nearest neighboursof each points and compute the first principal component of the corresponding matrix. This couldbe described in the following two steps,

1.1 Determine k nearest neighbors x_{ij} of x_i, j =1, · · ·,k, N matrices with $d*k$ size are obtained from this step.

1.2 Compute the first principal component$\mathbf{P}\in \mathbf{R}^{k\times 1}$ofthe N matrixes obtained from the step 1.1.

1.3 Compute the local orthogonal projection $\mathbf{O}\in \mathbf{R}^{k\times k}$.

$$\mathbf{O} = \mathbf{P}*\mathbf{P}'-\mathbf{I} \tag{1}$$

Where **I** is identity matrix of size $k*k$

1.4 Let I_i= {$i_{1,}$ ...,i_k}the set of indices forthe k nearest neighbors of x_i.Construct the square matrix **M**by locally summing the orthogonal projection based on the neighborhoods indices I_i:

$$\mathbf{M}(I_i, I_i) = \mathbf{M}(I_i, I_i)+\mathbf{O} \tag{2}$$

Step 2.Solve the eigenvalues and eigenvectors problem for theglobalmatrix **M**.

III. Datasets used

Several datasets have been used in this study for validation, error estimation and experiments. A Swiss-roll, which was created to test out various dimensionality reduction algorithms, has been used in this study for different purposes. It is generated randomly by sampling a 3D Swiss-roll surface with no class label information. The second one is the famous Iris data set provided by Anderson [12]. The data set has 4 features and 150 samples consisting of three species of Iris flower with 50 samples of each species.Microarray is another data set has been used in this study. The data is composed of 72 observations with 255 features. The 72 observations are divided into two clusters which separate individuals between the diseased (-1) and healthy (1).This data set has been pre-processed by applying a feature selection algorithm in order to remove the noise and irrelevant features which affect the result of dimensionality reduction algorithm [16].

IV. ERROR ESTIMATION of LPC

Different methods have been proposed for error estimation and quality measurement of dimensionality reduction. For this algorithm, we have used trustworthiness measurement proposed by Kaski et al [13] to measure the quality of the algorithm LPC. As our main issueof dimensionality reduction is to preserve the neighbourhoods of the points in the input space and output space, we have decided to use this type of quality measurement thatis based on the comparison of the neighbourhood of the points in the input and output space. For example if point x is close to points w and z in the input space**X**, then point x should be also close to points$w$ and z in the output reduced space**Y**otherwise an error arises after reduction. In Figure 1, we have a point X_i in the input space (left image) with the nearest neighbours points represented in red colour (point w and z). These points transformed and mapped into lower dimensional space. The point X_i transformed into another point Y_i. The red points still nearest neighbours for the point $Y_i$except the point *z* which becomes out of the nearest neighbours points. On the other hand, a blue point becomes a new nearest point for the point *Yi* whereit was not in the input space.In this case, we don't have a complete trustworthinessbecause the neighbourhood of the point Xi have been changed between the input and output space.

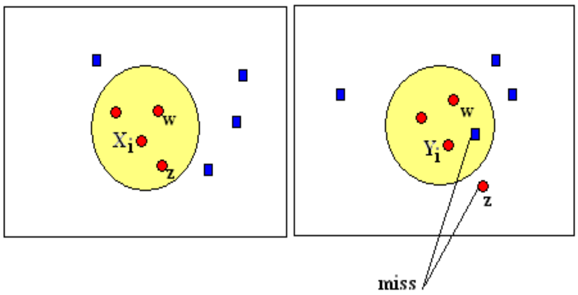

Fig. 1: Types of errors in reduction.

Trustworthiness aims to find to which extent neighbors in the input space also have corresponding neighbors in the output space by ranking of neighbourhood point sets in input and output space. The rule of trustworthiness can be defined as follows: Let *N* be the number of data samples and *r(i, j)* be the rank of the data sample *j* in the ordering according to the distance from *i* in the original data space. Denote by $U_k(i)$ the set of those data samples that are in the neighbourhood of size *k* of the sample *i* in the visualization display but not in the original data space [14]. The measure of trustworthiness $M_{trust}(k)$ of the dimensionality reduction is

$$M_{trust}(k) = 1 - A(k)\sum_{i=1}^{N}\sum_{j\in U_k(i)}(r(i,j)-k) \qquad (3)$$

Two data sets have been examined by trustworthiness; the first dataset is a Swiss-roll data set and the second one is microarray data set. The two data setshave been reduced several times with different values of the parameter *k*. Figure 2 and 3 show the obtained result of the trustworthiness of LPC applied on Swiss-roll data set and Microarray data set respectively. As can be seen from Figure 2, the trustworthiness is quite stable around the value of 0.98 for different values of parameter *k*.In Figure 3, the trustworthiness is dramatically changed based the parameter*k*, but it can be noticed that the trustworthiness has highestvalues for *k*>10and especially at *k*=14.

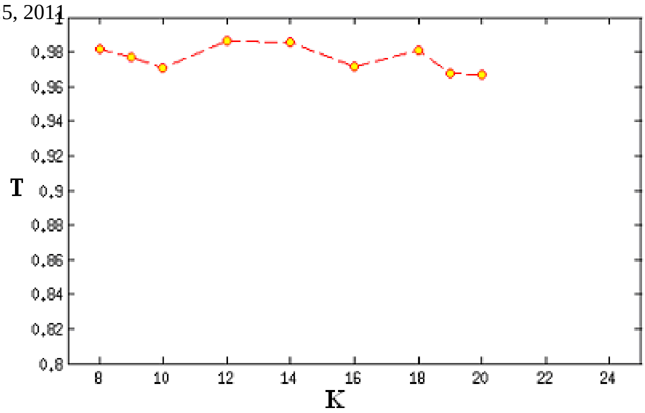

Fig. 2: Trustworthiness of LPC on Swiss-roll data set

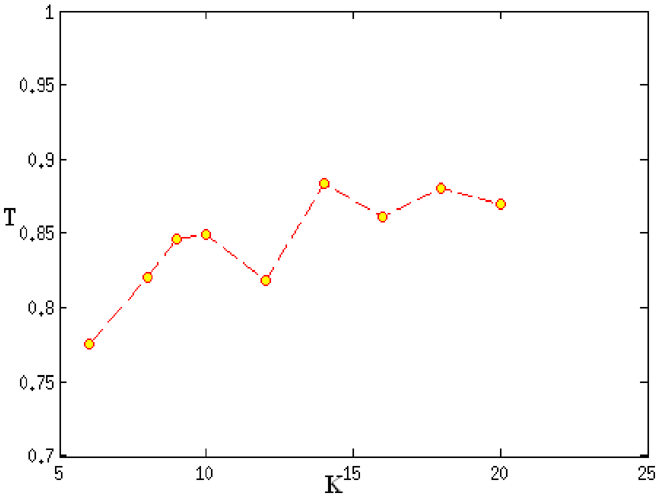

Fig. 3: Trustworthiness of LPC on Microarray data set

V. Experiments

In order to demonstrate the validity of the proposed algorithm, we performed experiments using Iris data set and artificial Swiss roll data set.

A. Validation Experiments

Iris data set: The algorithm is tested on theIris data set described in Section III.

Figure 4a represents the scattering of the original data set in 2D space. As can be seen, some data points from different classes are mixed together in the original 2D space. Figure 4b shows the data reduced to 2D data by LPC. In Figure 4b, the visualization performance shows that the three different classes are still separated even after the data has been reduced into two dimensions. Moreover, the trustworthiness measurement of this reduction is 0.995 for k=12. The number 0.995 means that the neighbourhoods of the data set are preserved with a very small error in the low dimensional space.

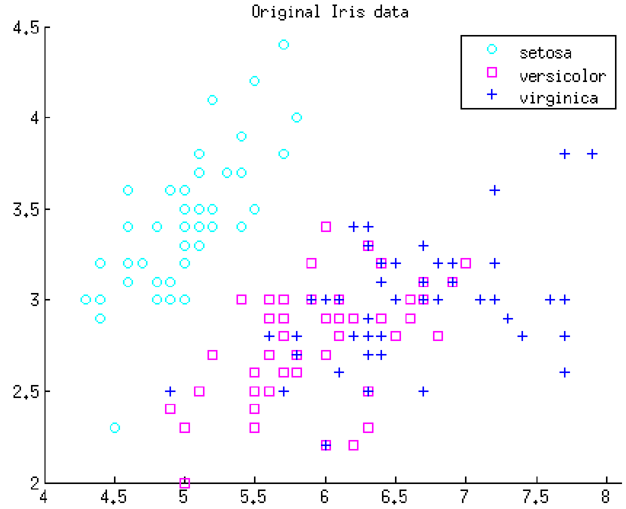

Fig. 4a: Original Iris data set

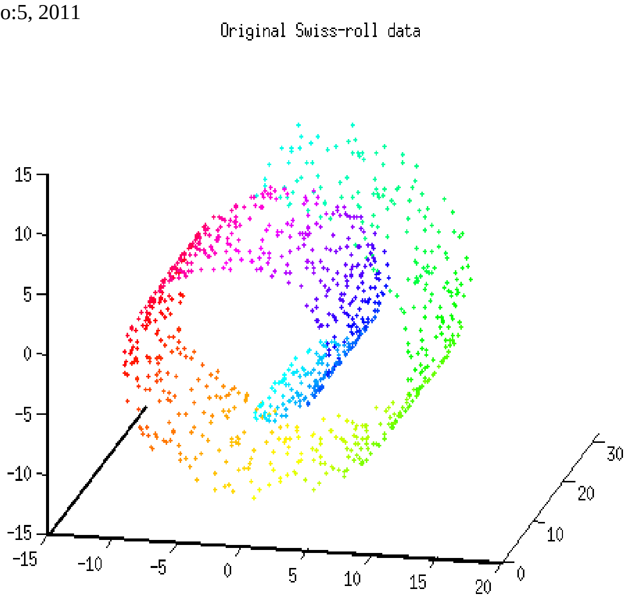

Fig. 5a: Swiss roll data set

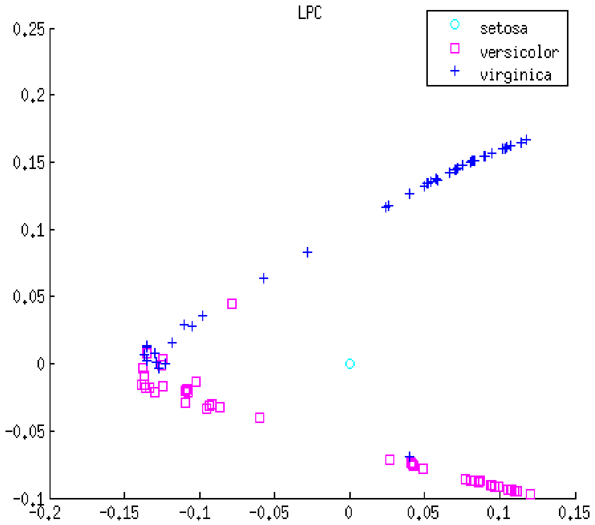

Fig. 4b: Iris data set processed by LPC

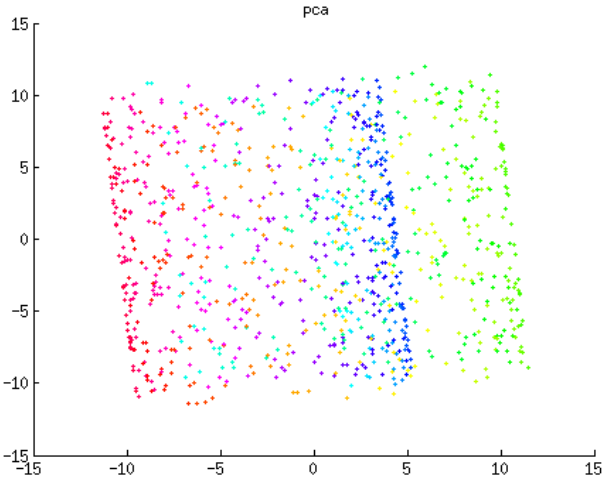

Fig. 5b: Swiss roll data reduced to 2D by PCA

Swiss Roll data set: As Roll data set was created to test out various dimensionality reduction algorithms, The algorithm is tested on the this data. In this experiment, we have generated 1000 points to test our algorithm.

Figure 5a represents the original data set in 3D space. As can be seen, the data are folded to have the Swiss-roll form. Figure 5b shows the data which reduced to 2D data by PCA is lacking the quality of visualization performance and dimensionality reduction. In Figure 5c, the data has better visualization than PCA and it shows that an adequate embedding preserving the shape of manifold can be achieved by LPC. In order to quantify the comparison between the two outputs, we have measured the trustworthiness of dimensionality reduction performed by PCA and LPC. The trustworthiness of LPC for Swiss-roll is 0.997 compared to 0.848 for PCA which suggests that LPC embedding is better than PCA for maintaining neighbourhood relationships.

B. Testing Experiments

*Microarray Data:*As our target from this algorithm is microarray data, a Leukaemia dataset has been used to demonstrate this algorithm.

The images below show the result of the obtained data set after applying PCA algorithm and LPC algorithm (Figure 6a and Figure 6b respectively).

It can be clearly seen that LPC reduced the data much better than PCA in terms of preserving the intrinsic dimensionality of the data. Also the trustworthiness of LPC is much better than PCA which has a value of 0.80. On the other hand, LPC has 0.86 value as a trustworthiness of preserving the neighbourhoods of the point in the low dimensional space.

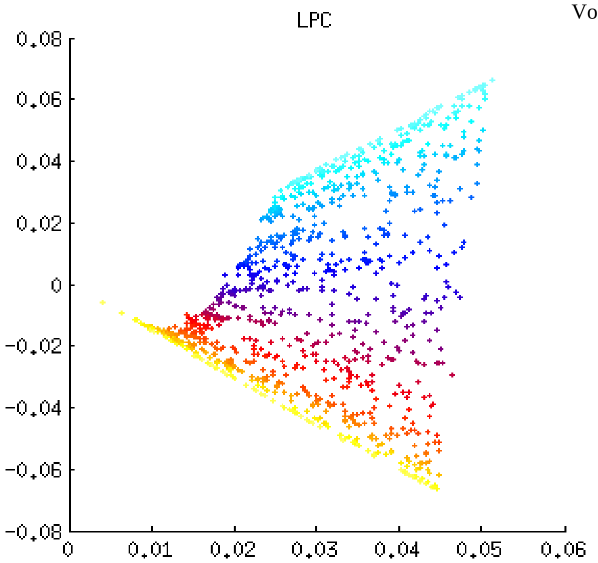

Fig. 5c: Swiss roll data reduced to 2D by LPC

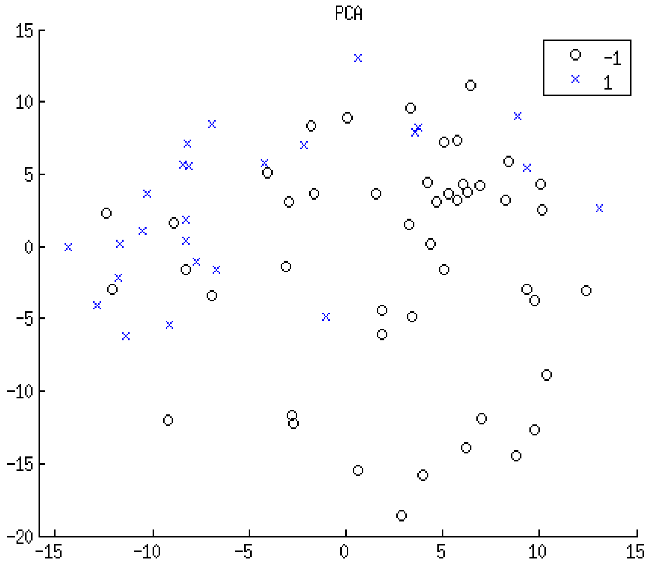

Fig. 6a: Leukaemia data reduced to 2D by PCA

VI. Comparison with other methods

Local principal component is a technique that is similar to locally embedding algorithms (LTSA [10] and LLE [2]) in that it constructs a local linear embedding of the *k* nearest neighbors. The idea of LPC is to have less instructions and computations with a good trustworthiness result because it is proposed for a complex microarray dataset. The algorithm of LPC aims to find local principal component around a data point x_i based on the *k* nearest neighbors of that point. This is followed by another step to extract the first principal component and then construct the square matrix from these principal components.

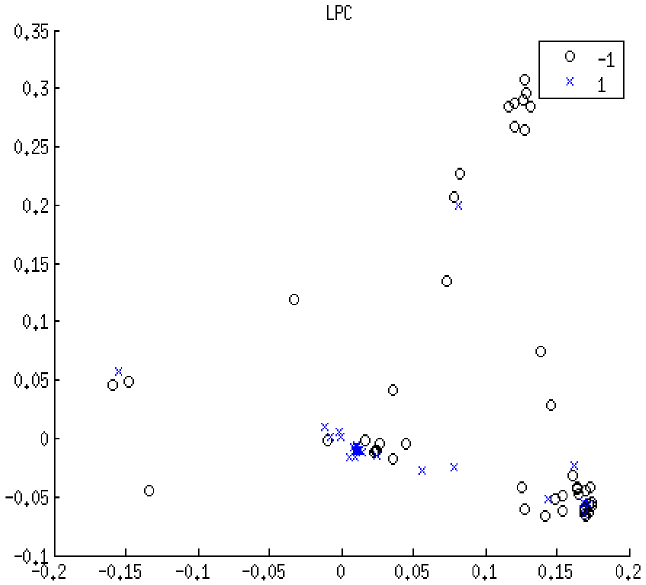

Fig. 6b: Leukaemia data reduced to 2D by LPC

Local Tangent Space Analysis (LTSA) is a technique that is similar to LPC as it describes local properties of the high-dimensional data using the local tangent space of each data point. However LPC has no restriction on the parameter *k* as in LTSA. On the other hand, LLE algorithm computes a different local quantity of the *k* nearest neighbors. Then, each data point is approximated with the best coefficients by a weighted linear combination of its neighbors in order to form a square matrix from these calculated values.

Several experiments have been done to make sure that LPC has good dimensionality reduction in terms of preserving the neighbourhoods of the points in the low dimensional space as in the high dimensional space. As our algorithm is similar to LLE, we present some experiment comparing LPC to LLE. Consequently, we have compared the trustworthiness of LPC to LLE applied on a Swiss-roll and microarray data set. Figure 4 and 5 present the result of these measurements.

As can be seen from the Figure 7 and 8, the trustworthiness of LPC is quite better than LLE at different values of the parameter *k* especially for *k*>10 for the microarray data. However, the trustworthinessof LPC is less than LLE just at k=6. With respect to the Swiss-roll comparison, the trustworthiness of LPC is quite stable around the value of 0.98 for different values of parameter *k* where the trustworthiness of LLE is in dramatically changes based on the parameter k. For example at *k*=14 the trustworthiness of LPC applied on Microarray data set is 0.89 which represent the maximum value. On the other hand, the trustworthiness of LLE is 0.85 at *k*=9 which represent maximum value as well.

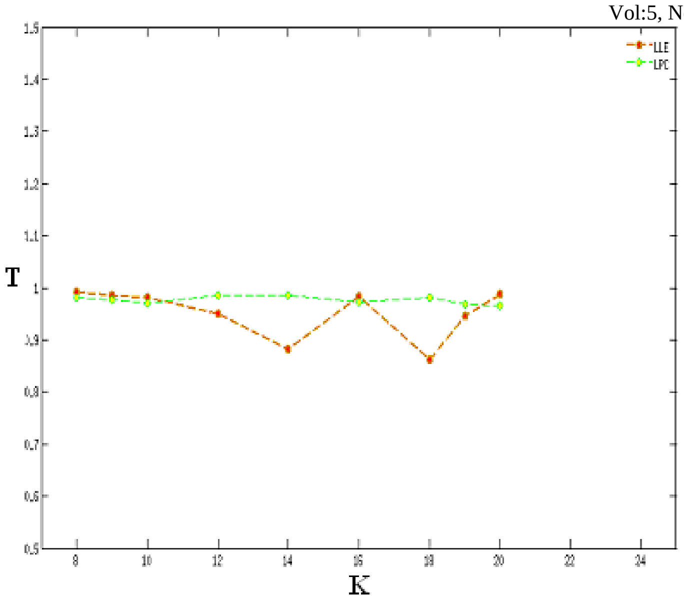

Fig. 7: Trustworthiness of LPC Vs LLE using Swiss-roll data set

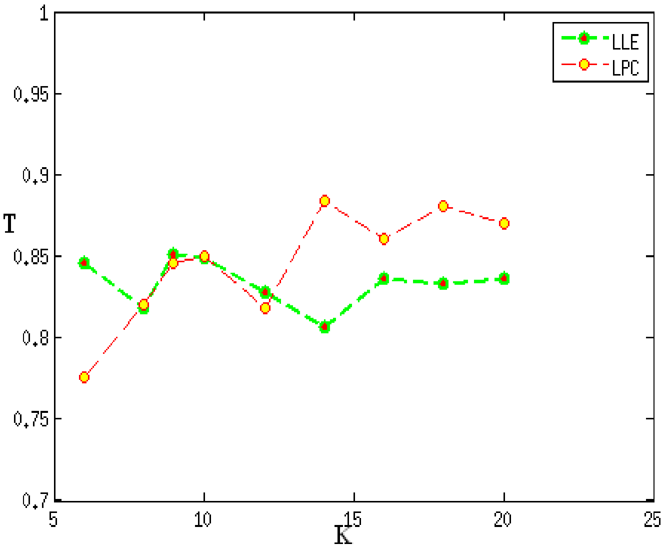

Fig. 8: Trustworthiness of LPC Vs LLE using Micro array data set

VII. Conclusion and Further Study

In this paper we have proposed an algorithm for high dimensional data reduction based on local principal component. We have discussed the validation experiment by applying the LPC on two different data sets (Iris and Swiss-roll). Moreover, we have applied LPC on a Leukaemia microarray data set. A good dimension reduction results have been demonstrated through these experiments and the algorithm outperform PCA in some aspects.

This algorithm provides a way to visualize data in order to see the position of a patient with respect to other patients. It also reduces high dimensional data into more easily handled low dimensional data.

Our future work is to make this algorithm as a supervised algorithm in order to have more accurate result. Another effective plan is to weight the features and include theses weights in the Euclidean distance measurement for retrieving the *k*nearest neighbours.

References

[1] V. Tenenbaum and J.C. Langford, A Global Geometric framework For Nonlinear Dimensionality reduction. Science, 290 (5500):23192323,2009.

[2] S.T. Roweis and L.K. Saul, Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science, 290(5500):23232326, 2000.

[3] C. Bowman, R. Baumgartner et al, Dimensionality Reduction for BiomedicalSpectra. Electrical and Computer Engineering, 2002. IEEE CCECE,2002.

[4] P. J. Kennedy, S. J. Simoff, D. Skillicorn and D. Catchpoole, Extracting and Explaining Biological Knowledge in Microarray Data. Proc. Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining, Sydney. (eds) Dai, H., Srikant, R., and Zhang, C., LNAI 3056, pp 699-703, Springer-Verlag Berlin, 2004.

[5] I. Guyon and A. Elisseeff, An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3 (2003) 1157-1182, 2002.

[6] J. Lee and M. Verleysen, Nonlinear Dimensionality Reduction Springer, 2007.

[7] J. Quansheng, J. Minping, et al., New approach of intelligent fault diagnosis based on LLE algorithm. Control and Decision Conference, 2008. CCDC 2008. Chinese, 2008.

[8] C. Varini, T. W. Nattkemper, et al., Breast MRI Data Analysis by LLE.Neural Networks, 2004.Proceedings. 2004 IEEE International Joint Conference, 2004.

[9] H. Tian, H. and D.G. Goodenough, Nonlinear Feature extraction of Hyperspectral Data Based on Locally Linear Embedding (LLE). In Geoscience and Remote Sensing Symposium, 2005.IGARSS '05.Proceedings.2005 IEEE International. 2005.

[10] Z. Zhang and H. Zha, Principal Manifolds and Nonlinear DimensionalityReduction Via Local tangent Space Alignment. SIAM Journal ofScientific Computing, 26(1):313338, 2004.

[11] D. Ridder and D. Rober, Locally Linear Embedding for classification. In the Pattern Recognition Group Technical Report Series. ICIP. 2005.

[12] E. Anderson, The Irises of the gasp Peninsula. Bulletin of the American Iris Society, 59(2-5), 1935.

[13] S. Kaski, J. Nikkila and et al., Trustworthiness and metrics in Visualizing Similarity of gene Expression. BMC Bioinformatics, 4:48, 2003.

[14] J. Venna, and S. Kaski, Visualizing gene Interaction Graphs With Local Multidimensional Scaling. In Michel Verleysen, editor, Proceedings of the 14th European Symposium on Artificial Neural Networks (ESANN2006), Bruges, Belgium, April 2628, pp. 557562, d-side, Evere, Belgium, 2006.

[15] K. Pearson, On Lines and Planes of Closest Fit to Systems of Points in Space . Philosophical Magazine, 2:559-572, 1901.

[16] A. Anaissi ,P. Kennedy and M. Goyal, A Framework for Very High Dimensional Data Reduction in the Microarray Domain . IEEE-BITA,2010.