LOCAL LINEAR PROJECTION (LLP)

Xiaoming Huo, Jihong Chen

School of Industrial & System Engineering, Georgia Institute of Technology
Atlanta, GA 30332-0205

ABSTRACT

Dimensionality reduction has important applications in exploratory data analysis. A method based on Local Linear Projection (LLP) is proposed. The advantage of this method is that it is robust against uncertainty. Statistical analysis is applied to estimate parameters. Simulational results on synthetic data are promising. Some preliminary experiment of applying this method to microarray data is reported. The results show that LLP can identify significant patterns. We propose some future tasks to perfect this method.

1. INTRODUCTION

Dimensionality reduction plays a significant role in exploratory data analysis. In many real applications, although the data may have very high dimensions, they typically embedded in manifolds (or subspaces) that are of substantially lower dimensions. Identifying these manifolds (or subspaces) are critical in understanding these data. It is also important in applications such as data visualization and modeling. In the communities of statistics, machine learning, and artificial intelligence, a substantial amount of techniques have been developed. In the following, we will give a quick review on works that are directly related to ours.

When the embedded structures are linear subspaces, linear techniques such as Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) can be used to identify the embedded linear subspaces. In PCA, the second order statistics (variances and covariances) of the data are considered, researchers find the directions in which the variances are maximized. SVD works on the data themselves. It finds the linear subspace that best preserves the information of the data. For both PCA and SVD, the embedded structure must be *globally* linear. In many applications, this condition is too restrictive. Multi-Dimensional Scaling (especially metric MDS) is close to PCA and SVD. PCA and SVD are to find the most significant linear subspaces. In Metric MDS, workers try to map the data into a low-dimensional space, at the same time keeping the inter-data distances [13]. Although the philosophical points are seemingly different, the underlying linear algebra are very similar.

When the global linearity condition is abandoned, some methods that focused on finding local embedded structures have been proposed, among them, we have for example principal curves [7, 2]. Recently, we have paid attention to some methods that are dedicated to identifying local hidden manifolds, for example, ISOMAP [11] and Local Linear Embedding (LLE) [8]. In ISOMAP, instead of consider the distance between two data points, they consider the geodesic distance, which is the length of the shortest path that resides on the embedded manifold. In implementations, this idea is realized by considering the k-nearest neighbors. Later on, in order to achieve better numerical performance, researchers have proposed some variations, e.g. Curvilinear Distance Analysis (CDA), [4]. In LLE, each data point is represented as a convex combination of its k-nearest neighbors; the data is then mapped into a low-D space, at the same time, the convex combinations (which is called embedding) is preserved to the best possibility. In [4, 11, 8], good examples are shown to illustrate these ideas. These examples are Swiss rolls, open boxes, and cylinders. We found them very instructive.

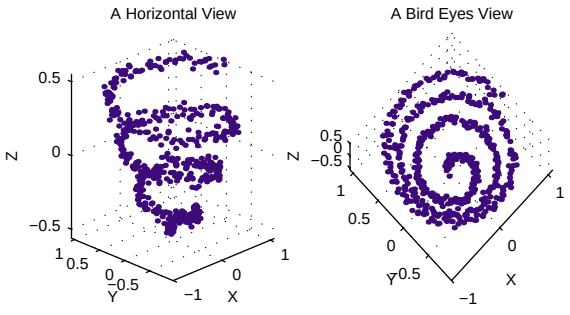

Fig. 1. Hurricane: A 3-D data with 1-D embedded structure.

In order to help our readers to visualize the type of the problem that we are trying to solve, we provide an exemplary data in Figure 1. This data is in 3-D but has an apparent 1-D embedded structure.

This work is partially supported by a seed grant from Center for Graphics Visualization and Usability at Georgia Institute of Technology and a DARPA-Lockheed-Martin-Stanford University contract.

Due to the maturization of the human Genome project and the availability of the microarray technology, microarray data poses a new challenge to data analysts. The microarray technology allows workers to measure the levels of gene expression for tens and thousands of genes simultaneously. The dimensionality of microarray data is definitely high. It is urgent to develop efficient dimension reduction tools. As a matter of fact, many previously mentioned tools have been applied to microarray data, for example, researchers have used SVD to interpolate missing values in a microarray data [12]. ISOMAP has been used to understand the structure of a microarray data [10]. PCA has been used to summarize microarray experiments [6]. A lot more examples can be found in the references of [5].

As an evidence to illustrate the importance of dimension reduction for microarray data, let us consider the clustering of genes. Clustering genes is to group together the genes that might be associated with identical functionalities. A nice survey on clustering methods for microarray datasets is given in [5]. An associated software is described in [9]. Many studies have been reported, e.g. [1]. Due to space, we can not enumerate all of them here. Dimension reduction can help improving the clustering result. One first project the data points to an embedded low-dimensional manifold, then compute the inter-distances between projections. The inter-distances should be more "faithful" than the inter-distance computed directly from the data. Hence a dimension reduction tool can be used as a preprocessing tool for a clustering algorithm.

A dimension reduction tool can also help to visualize the data. To visualize the data, we have to reduce the global dimensionality of the data. This is a little bit different from reducing the local dimensionality of a data. But by appending a post-processing method, it can be used to visualize the data. For example, we can look at the *local* structure of the data. In our simulational study to a synthetic data, we will give a demo of this idea.

In the works that we have seen so far, we observed the following shortcomings.

1. In many methods, (e.g. ISOMAP, CDA, LLE, and other k-nearest neighbor based methods,) no statistical model has been assumed. Hence it becomes difficult to quantitatively measure the success (or failure) of each method. It is also difficult to describe the domain in which these methods work.

2. Even though in most of the existing methods, the algorithms are clear and well described, while implementing them, there are always several parameters: for example, the number of nearest neighbors, and the dimension of the embedded manifold. No analysis on how to choose them have been fully reported.

We believe the answers to the above problems can be found through a statistical analysis, more specifically, the ANalysis Of VAriance (ANOVA).

In this paper, a statistical model is introduced to model the phenomenon of a locally embedded manifold in a *noisy* data. Based on the proposed model, we propose a Local Linear Projection method to identify this embedded manifold. Some preliminary computational and statistical analysis are carried out to determine how to choose the values of parameters in the model. We found that this method works well on synthetic data (as expected). We provide some preliminary results for microarray data.

The rest of the paper is organized as follows. In Section 2, the statistical model for embedded manifolds is described. In Section 3, we describe the idea and algorithm for LLP. In Section 4, some parameter estimation strategies are presented. In Section 5, we report simulational findings for both a synthetic data and a microarray data. In Section 6, questions that will be further analyzed are listed, and some final remarks are made.

2. MODEL

We assume an additive noise model. Suppose there are N observations, which are denoted by $y_1, y_2, \ldots, y_N$. Let p denote the dimension of each observation. We have $y_i \in \mathcal{R}^p, \forall 1 \leq i \leq N$. We assume that there is an underlying (piecewise smooth) function $f(\cdot)$ such that

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, 2, \ldots, N,$$

where variable $x_i \in \mathcal{R}^{p_0}$ is from a much lower dimensional space ($p_0 << p$), noises ε_i's follow a multivariate normal distribution ($\varepsilon_i \sim N(\vec{0}, \sigma^2 I_p)$, where σ is unknown).

In the above model, if the underlying function f is locally regular, or more specifically, function f can be approximated by a linear function:

$$f(x) \approx \beta_0 + \beta_1^T x,$$

where $\beta_0 \in \mathcal{R}^p$ and $\beta_0 \in \mathcal{R}^{p \times p_0}$, then locally linear projection can be applied to extract this information.

3. LOCAL LINEAR PROJECTION

LLP can be applied to extract the local low-dimensional structure. In the first step, neighboring observations are identified. In the second step, SVD or PCA is used to estimate the local linear subspace. Finally, the observation is projected into this subspace.

ALGORITHM: LLP

for each observation $y_i, i = 1, 2, 3, \ldots, N$,

1. Find the K-nearest neighbors of y_i. The neighboring points are denoted by $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_K$.
2. Use PCA or SVD to identify the linear subspace that contains most of the information on vectors $\tilde{y}_1$, $\tilde{y}_2$, $\ldots$, $\tilde{y}_K$. Suppose the linear subspace is $\mathcal{A}_i$, and $P_{\mathcal{A}_i}(x)$ denote the projection of a vector x into this subspace. Let k_0 denote the assumed dimension of the embedded manifold, then subspace $\mathcal{A}_i$ can be viewed as a linear subspace spanned by the vectors associated with the first k_0 singular values.
3. Project y_i into the linear subspace $\mathcal{A}_i$ and let $\widehat{y_i}$ denote this projection: $\widehat{y_i} = P_{\mathcal{A}_i}(y_i)$.

end.

The output of LLP, $\widehat{y_i}$, $i = 1, 2, \ldots, N$, are more "faithful" to the underlying structure (if it exists) than the original observations are.

A justification to step 2. is that based on the previous model, for $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_K$, we have

$$\begin{aligned} \tilde{y}_1 &= \beta_0 + \beta_1^T \tilde{x}_1 + \tilde{\varepsilon}_1, \\ \tilde{y}_2 &= \beta_0 + \beta_1^T \tilde{x}_2 + \tilde{\varepsilon}_2, \\ &\vdots \\ \tilde{y}_K &= \beta_0 + \beta_1^T \tilde{x}_K + \tilde{\varepsilon}_K, \end{aligned}$$

where $\tilde{\varepsilon}_i \sim N(\vec{0}, \sigma^2 I_p)$, and $\tilde{x}_i \in \mathcal{R}^{p_0}$, $i = 1, 2, \ldots, K$. Hence $\tilde{y}_i$'s can be viewed as random vectors whose mean vectors are from a low-dimensional subspace. The low-dimensional subspace can be extracted via SVD of vectors $\tilde{y}_1$, $\tilde{y}_2$, $\ldots$, $\tilde{y}_K$. The dimension of this linear subspace can be estimated by analyzing the variances.

The computational complexity of LLP is roughly $C(p, k_0, K)N^2$, where constant $C(p, k_0, K)$ is a function of the dimension of the data (p), the dimension of the embedded linear subspace (k_0), and the number of nearest neighbors (K). The reasoning is as follows. First of all, to identify the nearest neighbors, the distance matrix of the N observations need to be computed, which costs $O(pN^2)$ operations. Then each row (or column) of the distance matrix need to be sorted, which costs $O(N \log(N))$ (order of complexity for the quick sort algorithm) multiply with N (number of rows) operations. Or in other words, the sorting takes $O(N^2 \log(N))$ operations. Suppose that each SVD step takes a constant amount of operations $C_2(p, k_0, K)$, so does each projection step. Overall, the order of complexity for LLP is $C(p, k_0, K)N^2$.

4. ESTIMATING MODEL PARAMETERS

There are two key parameters in LLP. They are the number of nearest neighbors (K) and the dimension of the local underlying subspace (k_0). The ideal number for K is the one such that the linearity assumption holds. For parameter k_0, it is ideal to have $k_0 = p_0$.

4.1. Number of the Nearest Neighbors

Following the notations in Section 3, for a fixed data point y_i and its K-nearest neighbors $\tilde{y}_j$, $j = 1, 2, \ldots, K$, if the linearization model is true, the squared distances

$$d_{i,j} = \|y_i - \tilde{y}_j\|_2^2, \quad j = 1, 2, \ldots, K,$$

should approximately follow the $2\sigma^2 \cdot \chi_p^2$ distribution. These distances can be ordered:

$$d_{i,(1)} < d_{i,(2)} < \cdots < d_{i,(K)}.$$

If we calculate the differences

$$d_{i,(j+1)} - d_{i,(j)}, \quad j = 1, 2, \ldots, K-1,$$

we are going to observe a few big ones at the beginning, and then it decreases to small ones. This is because for χ^2-distributed random variables, the sequence of the differences of the order statistics is going to have the above mentioned pattern. The decreasing pattern of the differences can help to identify the appropriate number of nearest neighbors. Due to the space, we postpone detailed statistical analysis to future publications.

4.2. Dimension of the linear subspace

Still following the notations in Section 3, if a fatter version of the matrix β_1 is fixed, (in our case, it is computed from SVD,) then the analysis of the appropriate dimension of the linear subspace (p_0) falls into the domain of Analysis of Variance (ANOVA). In our case, the analysis is more complicated. Since the model matrix is computed from the data as well. Intuitively, as the dimension of the subspaces increases, people would expect a quick drop of variances at the beginning, and then a relatively steady decreasing. Again we postpone the detailed analysis to future publications.

5. SIMULATIONS

Two experiments are reported. The first one is for a synthetic signal. The second one is a microarray data which is also used in [3].

5.1. Synthetic Data: a 12-D Hurricane

A twelve dimensional signal is produced. The underlying function is

$$f(t) = (\sqrt{t}\sin(4.5\pi t),\ \sqrt{t}\cos(4.5\pi t),\ t - 0.5, \ldots)$$

The dimension $(4, 5, 6), (7, 8, 9)$, and $(10, 11, 12)$ has the same pattern as in the first three dimensions. This signal is intrinsically 1-D. An illustration of a noisy data that is generated based on the above function, limited to its first three dimensions, is in Figure 1. In creating the noisy data, the standard deviation for noise is chosen to be $\sigma = 0.10$. Based on the analysis of the differences between squared distances, we choose the number of the nearest neighbors $K = 40$. The results of applying LLP are shown in Figure 2.

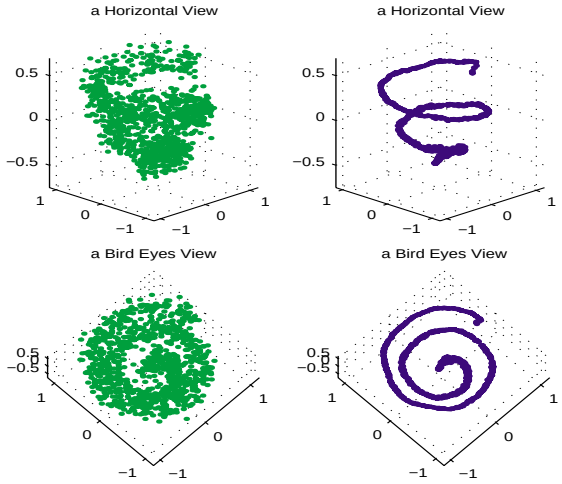

Fig. 2. Result of applying LLP to the 12-D synthetic data. Limited to the £rst three dimensions.

5.2. Microarray Data

The LLP is applied to a microarray dataset which is also used in [3]. (The dataset is downloadable on the web.) We found that the number of the nearest neighbors should be $K = 30$. The result of applying LLP is shown in Figure 3.

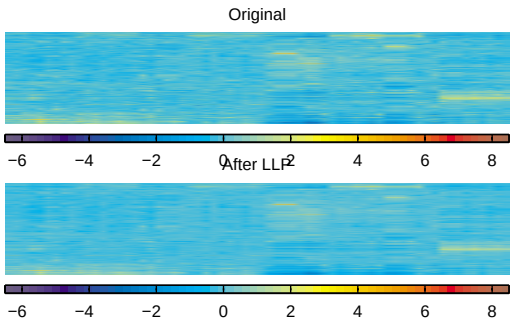

Fig. 3. Result of applying LLP to a microarray data. You must view it in a color figure.

We postpone the detailed discussion on this result to future publications.

6. FUTURE WORKS AND CONCLUSION

LLP has been useful in identifying locally low-dimensional embedded subspaces. It is an optimal dimension reduction tool. It can be used as a preprocessing tool for other data analysis techniques.

We plan to carry out a detailed analysis on the two approaches that are described in 4.1 and 4.2.

7. REFERENCES

[1] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns, *Proc. Natl Acad. Sci. USA*, vol 95: 14863-14868.

[2] Hastie, T., and Stuetzle, W. (1989) Principal Curves, *Journal of the American Statistical Association*, 84 (406): 502-516, June.

[3] Hastie, T., Tibshirani, R., and Friedman, J. H. (2001) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, *Springer series in statistics*, New York.

[4] Lee, J.A., Lendasse, A. and Verleysen, M. (2000) Curvilinear distance analysis versus isomap, *submitted to ESANN'02, Bruges*.

[5] Quackenbush, J. (2001) Computational analysis of microarray data, *Nat Rev Genet*, 6: 418-427.

[6] Raychaudhuri, S., Stuart, J.M. and Altman, R.B. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput. 2000*, 455-466.

[7] Stanford, D.C. and Raftery A.E. (2000) Finding curvilinear features in spatial point patterns: Principal curve clustering with noise, *IEEE Trans. PAMI*, 22 (6): 601-609, June.

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

[9] Sturn, A., Quackenbush, J. and Trajanoski, Z. (2002) Genesis: cluster analysis of microarray data, *Bioinfomatics*, 207-208.

[10] Tamayo, P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application, *Proc. Natl Acad. Sci. USA*, vol 96: 2907-2912.

[11] Tenenbaum, J.B., Silva, V., and Langford, J.C. (2000) A global geometric framework for nonlinear dimensionality reduction, *Science*, vol 290: 2319-2323.

[12] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tishirani, R., Bostein, D., and Altman, R.B. (2001) Missing value estimation methods for DNA microarrays, *Bioinformatics*, Vol.17(6): 520-525.

[13] Young, F. (1981) Introduction to Multidimensional Scaling: Theory, Methods, and Applications, *Academic Press*, New York.