Robust Nonlinear Dimensionality Reduction for Manifold Learning

Haifeng Chen   Guofei Jiang   Kenji Yoshihira
NEC Laboratories America, Inc.
4 Independence Way
Princeton, NJ 08540, USA
{haifeng, gfj, kenji}@nec-labs.com

Abstract

This paper proposes an effective preprocessing procedure for current manifold learning algorithms, such as LLE and ISOMAP, in order to make the reconstruction more robust to noise and outliers. Given a set of noisy data sampled from an underlying manifold, we first detect outliers by histogram analysis of the neighborhood distances of data points. The linear error-in-variables (EIV) model is then applied in each region to compute the locally smoothed values of data. Finally a number of locally smoothed values of each sample are combined together to obtain the global estimate of its noise-free coordinates. The fusion process is weighted by the fitness of EIV model in each region to account for the variation of curvatures of the manifold. Experimental results demonstrate that our preprocessing procedure enables the current manifold learning algorithms to achieve more robust and accurate reconstruction of nonlinear manifolds.

1. Introduction

Nonlinear manifold reconstruction has been drawing a surge of interest recently to find the best characterizations of high dimensional data. Current algorithms, such as local linear embedding(LLE) [8] and isometric feature mapping (ISOMAP) [9], are derived from the observation that even though the high dimensional data are nonlinear globally, they are often smooth and approximately linear in a local region. Hence the whole manifold is reconstructed based on the local geometry of each region. However it has been noted that these algorithms are sensitive to noises and outliers [1][2]. This is not desirable for real applications since the real-world data are often contaminated with noise and outliers due to the imperfect sensors or human mistakes. For that reason, this paper proposes a preprocessing procedure for current manifold learning algorithms in order to achieve a robust way of reconstructing the underlying nonlinear manifold.

Three steps are presented in the proposed procedure. First we analyze the neighborhood distances of each point to detect the outliers. A local smoothing step is then performed in each region based on the linear error-in-variables (EIV) modeling [10] of local structure. In the linear EIV model, *all* the points in the region are treated as noisy and their corresponding noise-free estimates are obtained from a numerically robust algorithm. Finally we propose a fusion based step to obtain the global estimate of noise-free coordinates for each sample from its several locally smoothed values. The fusion process is weighted by the fitness of linear model at each local region to account for the variation of curvatures of the underlying manifold. By doing so, the effects of high curvature regions are downplayed and more accurate estimate of noise-free coordinates are obtained.

Extensive experiments are performed to test the effectiveness of our proposed algorithm. The results based on both synthetic and real data illustrate that all the steps in the proposed procedure are helpful to reduce noises and identify outliers during the process of manifold reconstruction.

2. Proposed Approach

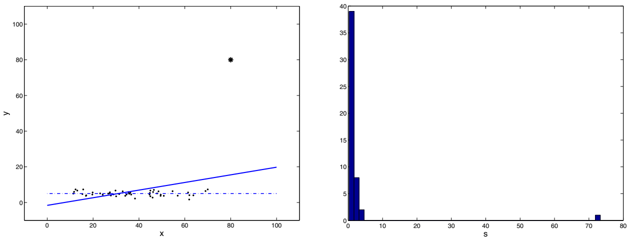
We begin with a set of sample points $\mathbf{Y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N]$, $\boldsymbol{y}_i \in R^D$, from a manifold $\mathcal{M}$ of intrinsic dimension $d < D$. The objective is to find the noise-free estimate $\tilde{\boldsymbol{y}}_i$ of each sample $\boldsymbol{y}_i$ as well as identify outliers if they exist. Three steps are presented in this section to achieve the goal.

2.1. Outlier Removal

Given the data set, the K nearest neighbors (K-NN) search is always performed first to define the local regions [8][9]. Many reconstruction algorithms, such as LLE, then apply least squares based linear fitting to discover the local structure of each region. However, a least squares based fitting algorithm is not robust to outliers. As shown in Figure 1(a), a single outlier will 'pull' the fitted line to an undesired direction. Although some robust regression tech-

0-7695-2521-0/06 $20.00 © 2006 IEEE

niques such as M-estimators and S-estimators exist in the statistics field [6], here we propose a more efficient and fast outlier detection method based on the data neighborhood information.



(a) (b)

Figure 1. (a)The effect of outliers; (b)The histogram of the s values.

The main observation for outlier detection is that the outliers are usually much *sparser* than inliers. That is, the distance between the outlier and its nearest patch is much greater than the distance between two near-by points in the patch. Therefore, we sort the distances $d(\boldsymbol{y}_i, \boldsymbol{y}_j)$ between each sample $\boldsymbol{y}_i$ and its neighbors $\boldsymbol{y}_j$s, which have already been calculated during the K-NN search, and define the lth smallest value

$$s_i = d_{[l]}(\boldsymbol{y}_i, \boldsymbol{y}_j) \qquad with \quad \boldsymbol{y}_j \in \mathcal{N}(\boldsymbol{y}_i) \tag{1}$$

where $\mathcal{N}(\boldsymbol{y}_i)$ represents the K nearest neighbors of $\boldsymbol{y}_i$. The choice of l depends on the knowledge of outlier distributions. Large value of l is for the situations that some outliers are clustered together to form a small island. Usually we choose $l = 2$.

As shown in the Figure 1(b), the s value of the outlier is much larger than those of inliers. Hence they can be identified based on the analysis of the histogram of s values. We calculate the robust mean of s_is by the least absolute deviation (LAD) location estimator [6], $\hat{u} = \underset{i}{\mathrm{med}}\, s_i$, and the robust scale by the median absolute deviations (MAD) [6] estimator $\hat{\sigma}_{MAD} = c \ \underset{i}{\mathrm{med}} |s_i - \underset{i}{\mathrm{med}}\, s_i|$ where $c = 1.4826$. Once we get the mean and scale estimate, the outlier is detected if its s value falls more than $4\hat{\sigma}_{MAD}$ far away from the mean.

2.2. Linear Error-in-Variables (EIV) Model

Now consider a set of D-dimensional vectors $\boldsymbol{y}_1^{(i)}, \cdots, \boldsymbol{y}_k^{(i)}$ located in the neighborhood of $\boldsymbol{y}_i$, where $k \leq K$ because of the removal of outliers. A local smoothing of those vectors is performed based on the geometry of local region that comprises the points. For simplicity, we use $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_k$ to denote those neighborhood points of $\boldsymbol{y}_i$.

In the LLE algorithm, the local geometry is represented by the weight vector, $\boldsymbol{w} = [w_{i1} \ \cdots \ w_{ik}]^\top$, that best reconstructs $\boldsymbol{y}_i$ from its neighbor points. By minimizing the following reconstruction error

$$\epsilon = \|\boldsymbol{y}_i - \sum_{j=1}^{k} w_{ij} \boldsymbol{y}_j\|^2 \tag{2}$$

subjected to $\sum_j w_{ij} = 1$, the LLE obtains the least squares solution of $\boldsymbol{w}$.

The equation (2) assumes that all the neighbors of $\boldsymbol{y}_i$ are free of noise, and only the observation $\boldsymbol{y}_i$ is noisy. This is frequently unrealistic since usually *all* the samples are corrupted by noise. As a result, the solution of (2) is biased. To remedy this problem, the error-in-variables (EIV) model is applied by minimizing

$$\epsilon' = \|\boldsymbol{y}_i - \sum_{j=1}^{k} w_{ij} \hat{\boldsymbol{y}}_j\|^2 + \sum_{j=1}^{k} \|\boldsymbol{y}_j - \hat{\boldsymbol{y}}_j\|^2 \tag{3}$$

subject to $\sum_j w_{ij} = 1$, where $\hat{\boldsymbol{y}}_j$ is the noise-free estimate of sample $\boldsymbol{y}_j$ in the local region surrounding the point $\boldsymbol{y}_i$.

The equation (3) can also be represented as

$$\epsilon' = \|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\|^2 + \sum_{j=1}^{k} \|\boldsymbol{y}_j - \hat{\boldsymbol{y}}_j\|^2 \tag{4}$$

taking into account that $\hat{\boldsymbol{y}}_i = \sum_j w_{ij} \hat{\boldsymbol{y}}_j$. Define the matrices $B = [\boldsymbol{y}_i \ \ \boldsymbol{y}_1 \ \ \boldsymbol{y}_2 \ \ \cdots \ \ \boldsymbol{y}_k] \in R^{D \times (k+1)}$ and $E = [\hat{\boldsymbol{y}}_i \ \ \hat{\boldsymbol{y}}_1 \ \ \hat{\boldsymbol{y}}_2 \ \ \cdots \ \ \hat{\boldsymbol{y}}_k] \in R^{D \times (k+1)}$ with $D \gg k$ for the common cases where the nonlinear manifold is embedded in a high dimensional space. The problem (4) can be reformulated as

$$\min \|B - E\|^2 \tag{5}$$

subject to

$$E\boldsymbol{\theta} = 0 \tag{6}$$

where $\boldsymbol{\theta} = [-1 \ \ w_{i1} \ \ w_{i2} \ \ \cdots \ \ w_{ik}]^\top$. From (6) the rank of E is k. Therefore the estimate E is the rank k approximation of the matrix B. If the singular value decomposition (SVD) of B is $B = \sum_{j=1}^{(k+1)} \sigma_j \mathbf{u}_j \mathbf{v}_j^\top$ with $\sigma_1 \geq \sigma_2 \cdots \geq \sigma_{k+1}$, we obtain the noise-free sample matrix E from the Eckart-Young-Mirsky Theorem [4] [7] as

$$E = \sum_{j=1}^{k} \sigma_j \mathbf{u}_j \mathbf{v}_j^\top \ . \tag{7}$$

The weight vector $\boldsymbol{w}$ can also be estimated from the SVD of matrix B. Since our purpose here is to find the local noise-free estimate E, we do not plan to discuss it in detail. For an in-depth treatment of EIV model and its solutions, please see [10]. According to [10], we can also obtain the first order approximation of the covariance $\hat{C}_w$ of the parameter $\boldsymbol{w}$, which is proportional to the estimation of the variance of noise

$$\hat{\sigma}_n^2 = \frac{\sigma_{k+1}^2}{D - k} \ . \tag{8}$$

2.3. Fusion

Since each sample $\boldsymbol{y}_i$ is usually included in the neighborhoods of other points as well as its own local region, it has several local noise-free estimates obtained from section 2.2. Given the set of different estimates, $\{\hat{\boldsymbol{y}}_i^{(1)}, \hat{\boldsymbol{y}}_i^{(2)}, \cdots, \hat{\boldsymbol{y}}_i^{(m)}\}$ with $m \geq 1$, our goal is to find the global noise-free estimate $\tilde{\boldsymbol{y}}_i$ of $\boldsymbol{y}_i$ from its many local values.

Due to the variation of curvatures of the underlying manifold, the linear model presented in section 2.2 may not always succeed in discovering local structures. In the local regions with large curvatures, the noise-free estimate $\hat{\boldsymbol{y}}_i^{(\cdot)}$ is not reliable. Suppose we have the covariance matrix C_j of $\hat{\boldsymbol{y}}_i^{(j)}$ to characterize the uncertainty of linear fitting through which $\hat{\boldsymbol{y}}_i^{(j)}$ was obtained, the global estimate of noise-free value $\tilde{\boldsymbol{y}}_i$ can be found by minimizing the sum of following Mahalanobis distances

$$\tilde{\boldsymbol{y}}_i = \underset{\tilde{y}_i}{\operatorname{argmin}} \sum_{j=1}^{m} (\tilde{\boldsymbol{y}}_i - \hat{\boldsymbol{y}}_i^{(j)})^\top \mathbf{C}_j^{-1} (\tilde{\boldsymbol{y}}_i - \hat{\boldsymbol{y}}_i^{(j)}) \,. \qquad (9)$$

The solution of (9) is

$$\tilde{\boldsymbol{y}}_i = \left(\sum_{j=1}^{m} \mathbf{C}_j^{-1} \right)^{-1} \sum_{j=1}^{m} \mathbf{C}_j^{-1} \hat{\boldsymbol{y}}_i^{(j)} \qquad (10)$$

i.e., the global estimate $\tilde{\boldsymbol{y}}_i$ is characterized by the covariance weighted average of the data. The more uncertain a local estimate is (the inverse of its covariance has a smaller norm), the less it contributes to the result of the global value.

The covariance matrix C_j can be calculated from the error propagation of the covariance of $\boldsymbol{w}$. However, since the dimension of $\hat{\boldsymbol{y}}_i^{(j)}$ is high, it is not feasible to use C_j directly. As an alternative we use the determinant of C_j to approximate (10)

$$\tilde{\boldsymbol{y}}_i = \left(\sum_{j=1}^{m} |\mathbf{C}_j|^{-1} \right)^{-1} \sum_{j=1}^{m} |\mathbf{C}_j|^{-1} \hat{\boldsymbol{y}}_i^{(j)} \qquad (11)$$

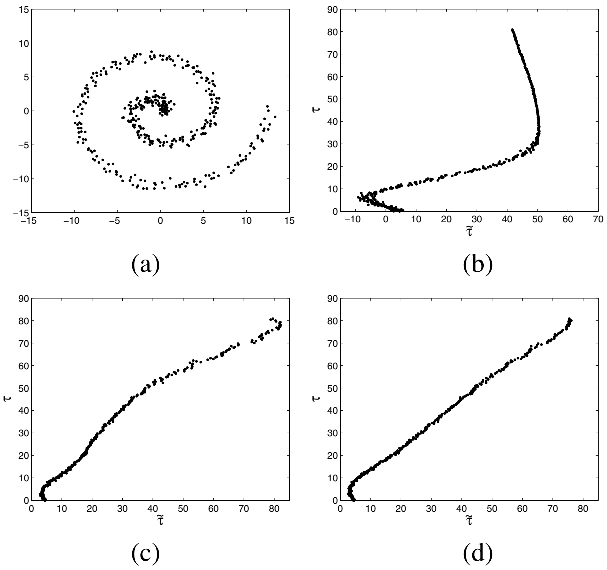
where $|C_j|$ is approximated by $|C_j| \approx \gamma \hat{\sigma}_n^2$. γ is a constant and does not contribute to the calculation (11).

3. Experimental Results

Several examples are presented in this section to demonstrate the effectiveness of our proposed procedure in noise reduction and outlier handling.

3.1. Noise Reduction

For the ease of visualization, an 1D manifold (curve) is used in this example. The 400 data points are generated by $g(t) = [t\cos(t), \;\; t\sin(t)]^\top$ added with certain amount of Gaussian noise, where t is uniformly sampled in the interval $[0, 4\pi]$. Figure 2(a) shows the curve under the noise with standard deviation $\eta = 0.3$. Since in this example the data dimension D(=2) is smaller than the size of neighborhood K (=12 for this dataset), we use a regularized solution of EIV model [5] to calculate the matrix E in (7).



(a) (b)

(c) (d)

Figure 2. (a)Samples from a noisy 1D manifold; (b)(c)(d)The centered arc length τ *vs.* manifold coordinates $\tilde{\tau}$ recovered by LLE, LLE with EIV modeling, LLE with EIV modeling and fusion respectively.

The reconstruction accuracy for 1D manifold is measured by the relationship between the recovered manifold coordinate $\tilde{\tau}$ and centered arc length $\tau(t)$ defined as $\tau(t) = \int_{t_0}^{t} \|J_g(t)\| dt$, where $J_g(t)$ is the Jacobian of $g(t)$, $J_g(t) = [\cos(t) - t\sin(t), \;\; \sin(t) + t\cos(t)]^\top$. The more accurate the manifold reconstruction is, the more linear is the relationship between $\tilde{\tau}$ and τ. Figure 2(b)(c)(d) shows their relationship curves generated by the LLE algorithm, LLE with EIV modeling, and LLE with both EIV modeling and fusion. It is obvious that the LLE with both EIV modeling and fusion performs better than the other two algorithms. Note in the LLE algorithm with only EIV modeling, the global noise-free estimate $\tilde{\boldsymbol{y}}_i$ is taken as the local smoothed value from the region that surrounds $\boldsymbol{y}_i$.

Further comparisons of the performance of three algorithms are carried out by some random simulations. The same manifold is used under different noise levels with standard deviation η from 0.1 to 1. At each noise level, 100 trials are run. We use the correlation coefficient between the recovered manifold coordinate $\tilde{\tau}$ and the centered arc length τ, $\rho = \frac{cov(\tilde{\tau},\tau)}{\sqrt{var(\tilde{\tau})var(\tau)}}$, to measure the strength of their linear relationship. Figure 3 shows the mean of correlation coefficient ρ, obtained by LLE, LLE with EIV modeling, and LLE with EIV modeling and fusion, for 100 trials under different noise levels with standard deviation η from 0.1 to 1. It illustrates that both the EIV modeling and fusion are beneficial to reducing the noise of data samples.

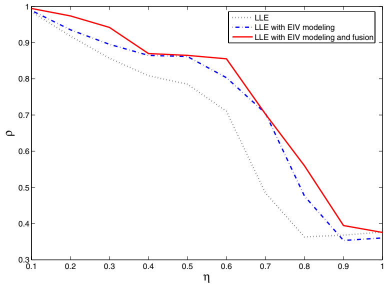


Figure 3. Performance comparison.

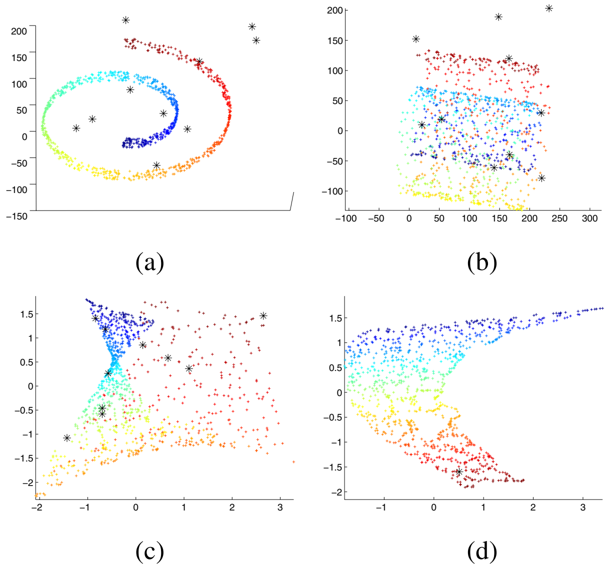
3.2. Outlier Detection

We generate 1000 3D data points $\boldsymbol{x}_i$ in which 990 are sampled from a 2D Swiss roll that is embedded in 3D space. The remaining 10 points are outliers, which are marked as '*' in Figure 4(a). We then transform the 3D data $\boldsymbol{x}_i$ into 100D vectors by an orthogonal transformation $\boldsymbol{y}_i = Q\boldsymbol{x}_i$, where $Q \in R^{100\times 3}$ is a random orthonormal matrix. Some Gaussian noises with standard deviation 0.5 are also added to the 100D vectors. Our outlier detector identifies 9 of the 10 outliers in this data set. Figure 4(d) shows the 2D coordinates computed by LLE with the outlier detection. Compared with the results of traditional linear methods such as PCA and the original manifold learning algorithms such as LLE, as shown Figure 4(a) and (b) respectively, our outlier detection plays an important role in reconstructing manifolds when the data set is corrupted with outliers.

We also apply our outlier identification approach to a real data set that is generated from a J2EE based web application. Each observation in that data set has 400 attributes, representing the number of interactions (calling relationships) between 20 different components in the application including Servlets and enterprise java beans (EJB). We generate 1500 observations during system normal operation, and 30 failure observations caused by different types of system faults such as *null call* and *expected exceptions* [3]. Our outlier detection procedure discovers 28 failure observations from the whole data set with only one false positive.

4. Conclusions

This paper has proposed an effective way of outlier handling and noise reduction for nonlinear manifold reconstruction. The proposed approach can serve as a preprocessing procedure for current algorithms in order to achieve more robust reconstructions of underlying manifolds. The experimental results have demonstrated the usefulness of proposed approach.



(a) (b)

(c) (d)

Figure 4. (a)Swiss roll with outliers; (b)(c)(d) 2D coordinates computed by PCA, LLE, and LLE with our outlier detection respectively.

References

[1] M. Balasubramanian and EL. Schwartz. The isomap algorithm and topological stability. *Science*, 295:7, 2005.

[2] M. Brand. Charting a manifold. In *Advances in Neural Information Processing Systems 15*. MIT Press, 2003.

[3] M. Chen, E. Kiciman, E. Fratkin, A. Fox, and E. Brewer. Pinpoint: Problem determination in large, dynamic systems. In *2002 International Performance and Dependability Symposium*, Washington, DC, June 2002.

[4] G. Eckart and G. Young. The approximation of one matrix by another of low rank. *Psychometrica*, 1:211–218, 1936.

[5] R.D. Fierro, G.H. Golub, P.C. Hansen, and D.P. O'Leary. Regularization by truncated total least squares. *SIAM Journal of Scientific Computing*, 18:1223–1241, 1997.

[6] P. J. Huber. *Robust Statistics*. New York:Wiley, first edition, 1996.

[7] L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *Quart. J. Math. Oxford*, 11:50–59, 1960.

[8] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[9] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[10] S. Van Huffel and J. Vandewalle. *The Total Least Squares Problem. Computational Aspects and Analysis*. Society for Industrial and Applied Mathematics, 1991.