Quality Assessment for Nonlinear Dimensionality Reduction using Procrustes Analysis

Erika Torres and Fu Dongmei

University of Science and Technology of Beijing,
Automation Science and Electrical Engineering School,
30 Xueyuanlu Beijing, 100083, P.R. China

Abstract. In order to achieve of PNL-ICA (Post-Nonlinear Independent Component Analysis) by using Dimensionality Reduction [1] the data set has to be embedded in a hyperplane, i.e. linear combination of the latent variables. This condition must be fulfilled in order to ensure that there is an unique solution to the problem. This article describes a new quality measure for Nonlinear Dimensionality Reduction based on Procrustes Analysis. This approach aims to solve the question of how to evaluate if a low-dimensional embedding (outcome of the process of dimensional reduction) can be used to recover data by using ICA methods.

Keywords: Nonlinear Dimensional Reduction, Post-Nonlinear Independent Component Analysis, Procrustes Analysis

1 Introduction

In the PNL-ICA case, the data set is in a manifold in $\mathbb{R}^{n}$, this manifold is in turn in a high-dimensional space with dimension $D > P$, this means that this data set does not fill the space completely. Hence, we can find a lower dimensional space ($\mathbb{R}^{n}$) where most of the variance of the data can be explained. It is valid to consider a subset of the data as a geometrical *Shape* with dimension $k \times D$, where k is the number of data points of the subset and D is dimension of the data before dimensional embedding. We can use the fact that every point of this subset is labeled, with preassigned correspondence between the original and the embedded data. Such data sets arise often in a biological or medical setting, when corresponding labeled points are called *landmarks*. Other areas in which landmark data arise include archaeology, astronomy, cartography, manufacturing, geology and geophysics. Thus in some instances landmarks may refer to the same physical markers identifiable in more than one map, satellite image, X-ray, etc.

2 Landmarks, Shape and Euclidean Distance

Suppose we have a post-nonlinear mixture of data $\mathbf{y}$, with $D \times N$ labeled points which can be embedded in a P-dimensional hyperplane $\hat{\mathbf{z}}$, this two sets- $\mathbf{y}$ and $\hat{\mathbf{z}}$- are closely related for two main reasons. First, because every data label in the P-dimensional space correspond to the same data point in the D-dimensional embedding (one-to-one relationship). Second, the information conveyed in the first set has to be coincident with the information conveyed in the second set. Since the model of PNL-ICA is isometric, then the key idea to obtain a satisfactory result is maintain the pair-wise geodesic distance when the data set is embedded into the low-dimensional space, if this part of the process is accurate, then the outcome will reduce the problem to a linear mixture version[1].

3 Geodesic Dissimilarity, a new quality measure

In order to measure the difference between geodesic distance (in the manifold) and Euclidean distance (in the hyperplane), first we have to break down the space in sufficiently small subsets, the so-called k-nearest neighbors (k-nn), in this part of the process we already estimate the Distance Matrix of $\mathbf{y}$ and $\hat{\mathbf{z}}$, D_{in} and D_{out} correspondingly. To find the k-nn, we sort the matrix $\mathbf{D}_{in}$, the index of elements of the sorted matrix are saved in a new matrix, then the first k elements of every column are selected, the other are ignored. With this information we can find the k-nn of the point i of D_{in} in D_{out}, with this two sets of $k \cdot k$ of distances we apply Procrustes shape distance [2], in every data point's k-nn in the two sets. After this process we calculate the total error adding every iteration result. The minimum Geodesic dissimilarity between two sets is zero and the maximum is 2, thus the best result for PNL-ICA would be those that have a Geodesic dissimilarity very near to zero.

The Geodesic Dissimilarity has two key aspects, uses the distance matrices to find the difference between the geodesic distance in the manifold and the Euclidean distance in the hyperplane and uses the same equation from Procrustes distance, with the difference that the points to be compared are the distance matrices, not the data sets. This new approach also uses k-nearest neighbors to compare the distances in small locally linear patches, this same idea has been successfully applied in a number of nonlinear dimensionality reduction algorithms.

In the practical case we do not have a linear mixture of the data to find the goodness of fit. Hence, to be able to use this quality measure in a nonlinear manifold, we apply the Procrustes distance over every i data point's k-neighborhood.

Algorithm 1 Geodesic Dissimilarity measure algorithm

procedure GEODISS($\mathbf{y}$, $\hat{\mathbf{z}}$, k, N) . N:# samples, k:# nearest neighbors
  Calculate the Distance matrices of $\mathbf{y}$ and $\hat{\mathbf{z}}$, D_{in} and $\mathbf{D}_{out}$.
  for $i \leftarrow 1, N$ **do**
    Choose the k-nn of the i-th point in D_{out}, save this distances in the matrix $\mathbf{X}$.
    Choose the same k-nn from the step 4 but in D_{in}, save this distances in the matrix $\mathbf{Y}$
    Calculate Procrustes shape distance between $\mathbf{X}$ and $\mathbf{Y}$.
    Save the result in the vector d_p.
    $i \leftarrow i + 1$.
  $G_d = \prod_{i} d_p$.
  end for
  return G_d.
end procedure

4 Experiments

We are going to use two basic signals represented by the data vector $\mathbf{x}$ in Equation 1. This signals are artificially mixed in a Post-Nonlinear way as follows in Equation 1:

$$\mathbf{x} = \begin{matrix} \sim \arccos(\cos(0.034\pi t)) \sim \\ \sin(0.006\pi t) \end{matrix}, \mathbf{z} = \begin{matrix} \sim 0.1\ 0.9 \sim \\ 0.8\ 0.2 \end{matrix} \mathbf{x}, \mathbf{y} = 0.5 \leq \begin{matrix} \Upsilon \\ +\square 2 + 1 + 1 \\ \square\square 2 + 1 + 1 \\ +0\ \square\square 2 + \square 2 \end{matrix} \begin{matrix} / \Upsilon \\ \\ f \\ \\ \end{matrix} \begin{matrix} / \\ \cos(\pi z_1) \\ \sin(\pi z_1) \\ \pi z_2 \end{matrix} f \quad (1)$$

We run the experiment explained above using 4 NLDR algorithms, which were extracted from *MANIfold Learning Matlab Demo* in http://www.math.ucla.edu/~wittman/mani/. We compared the goodness of fit evaluated with Trustworthiness, Continuity [3] and Geodesic Dissimilarity, the results are shown in Table 1. In this table, we can clearly see that Trustworthiness is an quite insensitive to evaluate the task of interest. Geodesic Dissimilarity can be used to compare the performance with different NLDR algorithms in order to choose the embedding that is most effective, the one with the lowest Geodesic Dissimilarity.

Algorithm	Trustworthiness	Continuity	Geodesic Dissimilarity	Is effective?
Locally Linear Embedding[4]	0.9991	0.9991	0.0710	YES
Curvilinear Component Analysis[5]	0.9765	0.9781	0.3512	NO
ISOMAP[6]	0.9999	0.9999	0.0013	YES
Laplacian Eigenmaps[7]	0.9972	0.9978	0.1754	NO

Table 1. Trustworthiness,Continuity and Geodesic Dissimilarity for the experiment described in the Section 4.

5 CONCLUSIONS

The disadvantage of the measures that are based on all pairwise distances is that they are not directly related to any specific task. Geodesic Dissimilarity offers an specific quality measure for dimensionality algorithms that embed the nonlinear manifold into a hyperplane and is directly related to the task where the user is trying to separate latent variables using ICA. This approach has been tested with speech signals with good results, with several experiments we can suggest that a embedding with Geodesic Dissimilarity less than 0.1 is effective to be used with ICA methods. An interesting fact, is that even non isometric NLDR algorithms also perform well for PNL-ICA, a deeper study of this phenomenon is left as future work. An important advantage of this quality measure is that is no directly related to visualization, is more related to the linear correlation between the data in the two embeddings, which can be useful in a number of applications, not only in PNL-ICA.

References

1. Lee, J.A., Jutten, C., Verleysen, M.: Non-linear ica by using isometric dimensionality reduction. ICA 2004 LNCS 3195 (2004) 710717
2. Gower, J.C.: Generalized procrustes analysis. Psychometrika Vol 40, No. 1 (1975) 33–51
3. Venna, J., Kaski, S.: Neighborhood preservation in nonlinear projection methods: An experimental study. In: Proceedings of the International Conference on Artificial Neural Networks. ICANN '01, London, UK, Springer-Verlag (2001) 485–491
4. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290** (2000) 2323–2326
5. Demartines, P., Herault, J.: Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. IEEE Transactions on Neural Networks 8 No. 1 (1997) 148–154
6. Tenembaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290** (2000) 2319–2323
7. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15 (2002) 1373–1396