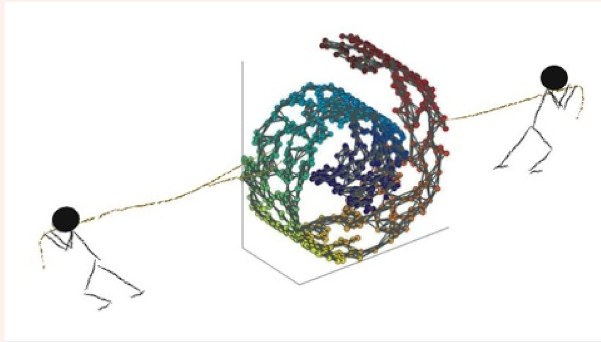
Non-Linear Dimensionality Reduction

YU.WU @liv.ac.uk

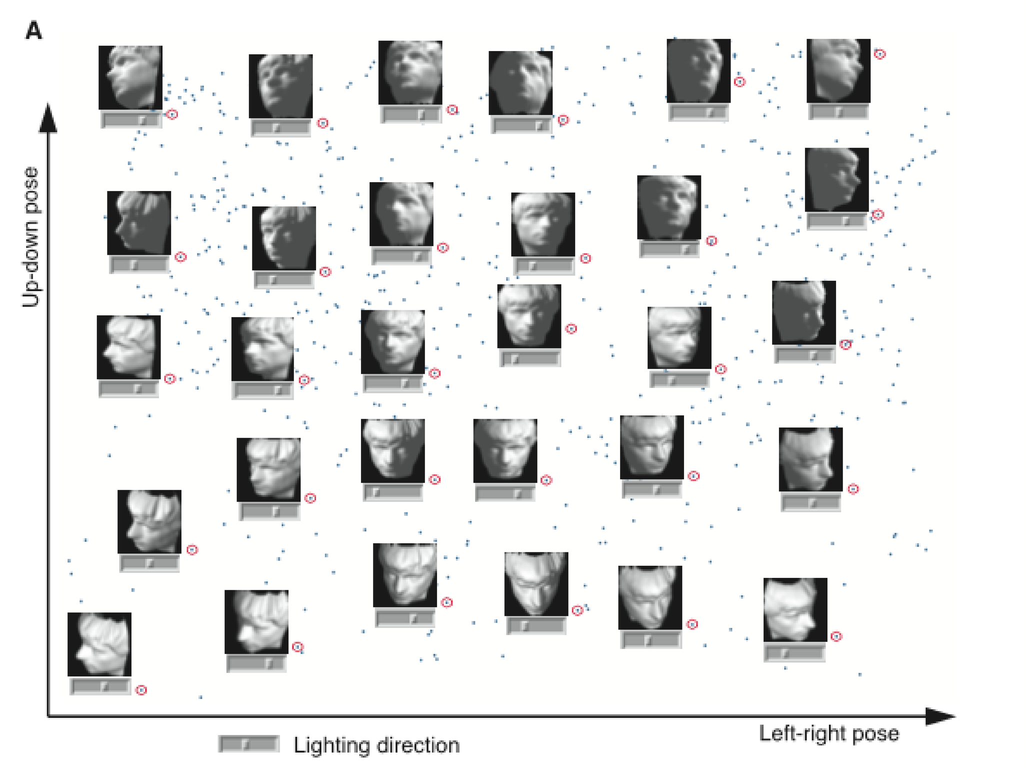


Abstract

- Scientists working with large volumes of high-dimensional data, such as text/linguistic data, video/images, global climate patterns, stellar spectra, or human gene distributions, regularly confronted the problem to *embed* data that originally lies in a high dimensional space in a lower dimensional space.
- Because for any high dimensional data to be interesting, it must perceive some intrinsic low dimensional nature, which makes dimensionality reduction possible.
- Laplacian Eigenmaps(LE)[1] is therefore introduced as one of the many Non-Linear Dimensionality Reduction(NLDR) techniques and various *NLDR* algorithms have been represented and compared.

Introduction

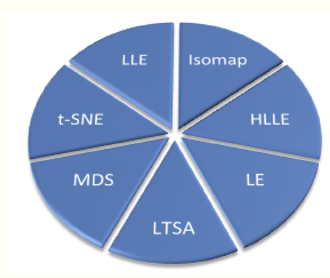
A canonical problem in *non-linear dimensional reduction* is illustrated in the following.



Embedding of High Dimensional Images

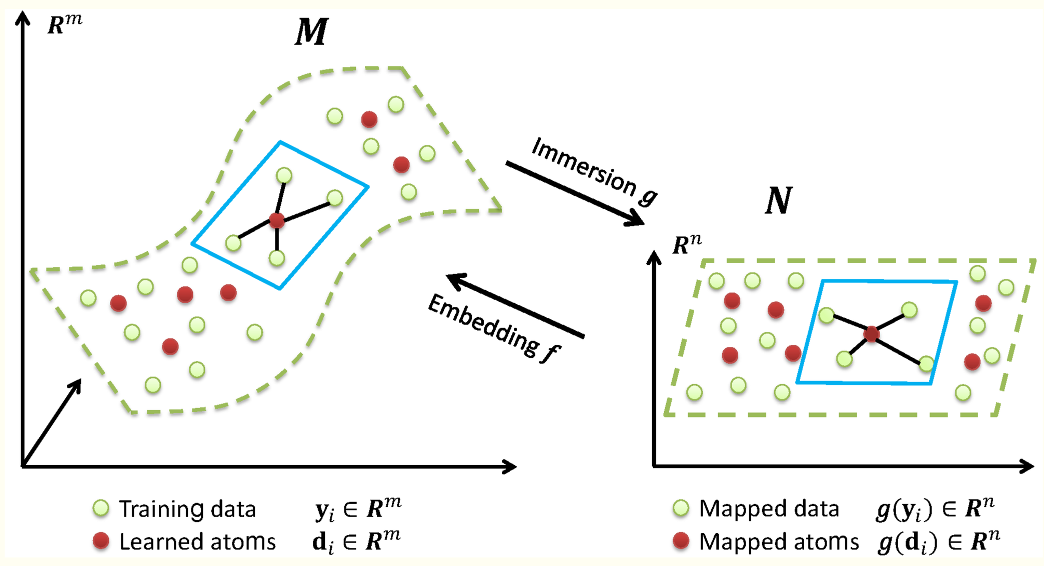
1. Many images of a person's face have been observed under different pose and lighting conditions, in no particular order.
2. Although the input dimensionality may be quite high (e.g., 4096 for these 64 pixel by 64 pixel images), the perceptually meaningful structure of these images has many fewer independent degrees of freedom.
3. All images lie on an intrinsically *three-dimensional* manifold (see the above figure), or constraint surface, that can be parameterised by two pose variables plus an azimuthal lighting angle.

Methods



In the following, seven NDR algorithms in this pie chart have been experimented on one digits datasets. Among these, we have studied the formulation of Laplacian Eigenmaps (also referred as Spectral Embedding) algorithms .

Laplacian Eigenmaps



Considering the problem of representing all of the vectors in a set of n d-dimensional samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ by a single vector $\boldsymbol{y} = y_1, y_2, \ldots, y_n$ such that y_i represents $\boldsymbol{x}_i$.

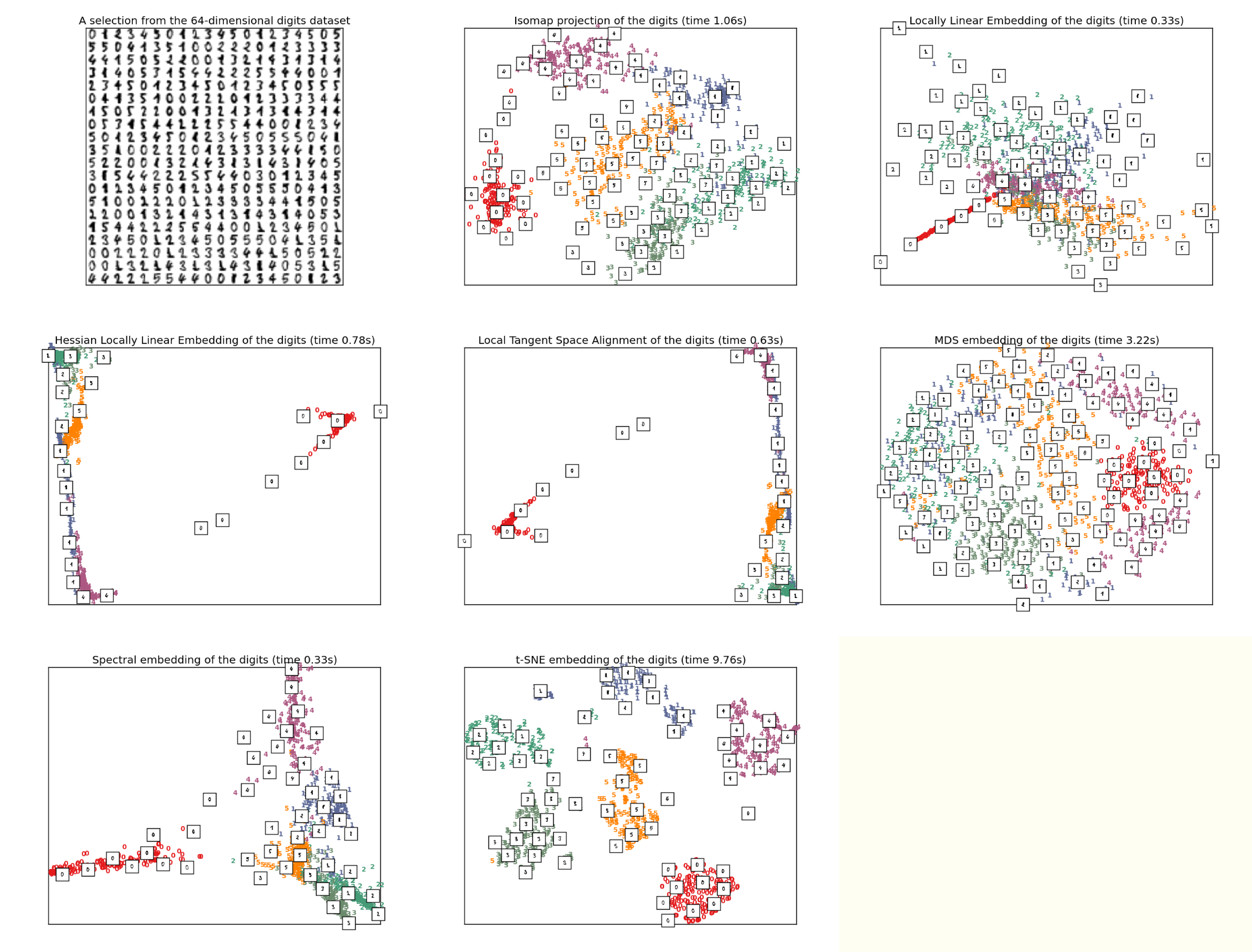
Laplacian Eigenmaps(also referred as Spectral Embedding) defines a "good" map by minimising the following objective functions under appropriate constraints,

$$\sum_{i,j}^{n} (y_i - y_j)^2 W_{ij},$$

where matrix $\mathbf{W} = [W_{ij}]$ is a *similarity matrix*.

This objective functions with the choice of symmetric weights $W_{ij}(W_{ij} = W_{ji})$ incurs a heavy penalty if neighbouring points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are mapped far apart, i.e, if $(y_i - y_j)^2$ is large. Therefore, minimising it is an attempt to ensure that, if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are "close", then y_i and y_j are close as well.

Results



Comparison of Various NDR Algorithms Upon A Selection of Digits Datasets

A Future Direction

NLDR algorithms also use pairwise *similarity matrix* as input, thus, becomes potential for developing relational learning algorithms[1][2], which aims at handling huge amounts of structured data generated daily in many application domains ranging from computational biology or information retrieval, to natural language processing.

Source Code

The source code for all algorithms are obtained at

`http://scikit-learn.org/stable/modules/manifold.html`

References

[1] Belkin, Mikhail and Niyogi, Partha. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. In *Neural Computation*

[2] Van der Maaten, Laurens, and Geoffrey Hinton. Visualizing Data Using t-SNE. In *JMLR*

[3] Bordes, Antoine, et al. A Semantic Matching Energy Function for Learning With Multi-relational Data. In *Machine Learning*