Grouping and dimensionality reduction by locally linear embedding

Marzia Polito  
Division of Physics, Mathematics and Astronomy  
California Institute of Technology  
Pasadena, CA, 91125  
polito@caltech.edu

Pietro Perona  
Division of Engineering and Applied Mathematics  
California Institute of Technology  
Pasadena, CA, 91125  
perona@caltech.edu

Abstract

Locally Linear Embedding (LLE) is an elegant nonlinear dimensionality-reduction technique recently introduced by Roweis and Saul [2]. It fails when the data is divided into separate groups. We study a variant of LLE that can simultaneously group the data and calculate local embedding of each group. An estimate for the upper bound on the intrinsic dimension of the data set is obtained automatically.

1 Introduction

Consider a collection of $N$ data points $X_i \in \mathbb{R}^D$. Suppose that, while the dimension $D$ is large, we have independent information suggesting that the data are distributed on a manifold of dimension $d \ll D$. In many circumstances it is beneficial to calculate the coordinates $Y_i \in \mathbb{R}^d$ of the data on the lower-dimensional manifold, both because the shape of the manifold may yield some insight in the process that produced the data, and because it is cheaper to store and manipulate the data when it is embedded in fewer dimensions. How can we compute such coordinates?

Principal component analysis (PCA) is a classical technique which works well when the data lie close to a flat manifold [1]. Elegant methods for dealing with data that is distributed on curved manifolds have been recently proposed [3, 2]. We study one of them, Locally Linear Embedding (LLE) [2], by Roweis and Saul. While LLE is not designed to handle data that are disconnected, i.e. separated into groups, we show that a simple variation of the method will handle this situation correctly. Furthermore, both the number of groups and the upper bound on the intrinsic dimension of the data may be estimated automatically, rather than being given a-priori.
2 Locally linear embedding

The key insight inspiring LLE is that, while the data may not lie close to a "globally linear" manifold, it may be "approximately locally linear", and in this case each point may be approximated as a linear combination of its nearest neighbors. The coefficients of this linear combination carry the vital information for constructing a lower-dimensional linear embedding.

More explicitly, consider a data set \( \{X_i\}_{i=1}^{N} \in \mathbb{R}^D \). The local linear structure can be easily encoded in a sparse \( N \times N \) matrix \( W \), proceeding as follows.

The first step is to choose a criterion to determine the neighbors of each point. Roweis and Saul chose an integer number \( K \) and pick, for every point, the \( K \) points nearest to it. For each point \( X_i \) then, they determine the linear combination of its neighbors which best approximates the point itself. The coefficients of such linear combinations are computed by minimizing the quadratic cost function:

\[
\varepsilon(W) = \sum_{i} |X_i - \sum_{j=1}^{N} W_{ij} X_j|^2
\]

while enforcing the constraints \( \sum_{j=1}^{N} W_{ij} = 0 \) if \( X_j \) is not a neighbor of \( X_i \), and \( \sum_{j=1}^{N} W_{ij} = 1 \) for every \( i \); these constraints ensure that the approximation of \( X_i \approx X_i = \sum_{j=1}^{N} W_{ij} X_j \) lies in the affine subspace generated by the \( K \) nearest neighbors of \( X_i \), and that the solution \( W \) is translation-invariant. This least square problem may be solved in closed form [2].

The next step consists of calculating a set \( \{Y_i\}_{i=1}^{N} \) of points in \( \mathbb{R}^d \), reproducing as faithfully as possible the local linear structure encoded in \( W \). This is done minimizing a cost function

\[
\Phi(Y) = \sum_{i=1}^{N} |Y_i - \sum_{j=1}^{N} W_{ij} Y_j|^2
\]

To ensure the uniqueness of the solution two constraint are imposed: translation invariance by placing the center of gravity of the data in the origin, i.e. \( \sum Y_i = 0 \), and normalized unit covariance of the \( Y_i \)'s, i.e. \( \frac{1}{N} \sum_{i=1}^{N} Y_i \otimes Y_i = I \).

Roweis and Saul prove that \( \Phi(Y) = \text{tr}(Y^T MY) \), where \( M \) is defined as

\[
M = (I - W)^T(I-W)
\]

The minimum of the function \( \Phi(Y) \) for the \( d \)-th dimensional representation is then obtained with the following recipe. Given \( d \), consider the \( d+1 \) eigenvectors associated to the \( d+1 \) smallest eigenvalues of the matrix \( M \). Then discard the very first one. The rows of the matrix \( Y \) whose columns are given by such \( d \) eigenvectors give the desired solution. The first eigenvector is discarded because it is a vector composed of all ones, with 0 as eigenvalue. As we shall see, this is true when the data set is 'connected'.

2.1 Disjoint components

In LLE every data point has a set of \( K \) neighbors. This allows us to partition of the whole data set \( X \) into \( K \)-connected components, corresponding to the intuitive visual notion of different 'groups' in the data set.

We say that a partition \( X = \cup_i U_i \) is finer than a partition \( X = \cup_j V_j \) if every \( U_i \) is contained in some \( V_j \). The partition in \( K \)-connected components is the finest