Abstract

We propose a new dimensionality reduction method, the elastic embedding (EE), that optimises an intuitive, nonlinear objective function of the low-dimensional coordinates of the data. The method reveals a fundamental relation between a spectral method, Laplacian eigenmaps, and a nonlinear method, stochastic neighbour embedding; and shows that EE can be seen as learning both the coordinates and the affinities between data points. We give a homotopy method to train EE, characterising the critical value of the homotopy parameter, and study the method's behaviour. For a fixed homotopy parameter, we give a globally convergent iterative algorithm that is very effective and requires no user parameters. Finally, we give an extension to out-of-sample points. In standard datasets, EE obtains results as good or better than SNE's, but more efficiently and robustly.

2 A relation between Laplacian eigenmaps and stochastic neighbor embedding

Laplacian eigenmaps (LE): given affinities \( W \) for data points \( y_1, \ldots, y_N \),

\[
\min \sum_{i,j} w_{ij} (\|x_i - x_j\|^2) - \lambda \sum_{i,j} w_{ij} \exp(-\|x_i - x_j\|^2)
\]

s.t. translation & scale constraints on \( x \).

Dislikes placing far apart latent points that correspond to similar data points, but places no direct constraint on pairs associated with distant data points.

Often leads to distorted maps: large clusters of points collapse, local clusters and gaps, boundary effects.

Global optimum from spectral problem: eigenvectors of \( L = D - W \) with \( D = \text{diag}(\sum_i w_{ij}) \).

Stochastic neighbor embedding (SNE):\

\[
\sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}
\]

\( p_{ij} = \text{softmax}(d_{ij}) \), \( q_{ij} = \text{softmax}(d_{ij}) \).

The repulsion in the prior term in \( E \) drops off rapidly, so the embedding \( X \) has a characteristic scale at each \( \lambda \) that increases with \( \lambda \).

The embedding \( X(\lambda) \) undergoes a series of bifurcations where an eigenvalue of the Hessian of \( E \) becomes negative. At some of these \( \lambda \), \( X \) increases.

At the first bifurcation \( \lambda = \lambda^* \), \( X \) grows from \( \lambda = 0 \) to \( \lambda^* \), expanding along the trailing eigenvector of \( L - \lambda I \) (very similar to a 1 D LE embedding).

We have upper and lower bounds for \( \lambda^* \) (see paper).

As \( \lambda \) increases, \( X \) unfolds globally and reorganises locally, representing better the global and local manifold structure. This is the region of \( \lambda \) where the best embeddings occur (region 0 in the Swiss roll).

For large \( \lambda \), the points distribute approximately equidistant from each other locally (first grid in 2D) while maintaining the global structure of the manifold, and the map scale increases logarithmically with \( \lambda \).

The learned affinities start as Gaussian for small \( \lambda \) and become Mexican-hat functions as \( \lambda \) increases.

We illustrate this for a Swiss roll dataset with \( N = 2000 \) points in \( D = 3 \).

Squared diameter of \( X(\lambda) \) and illustrative stages of map development:

\begin{itemize}
  \item \( d \approx 10^{-4} \): Gaussian
  \item \( d \approx 1 \): Mexican hat
\end{itemize}

Learned affinities \( w_{ij} = \text{softmax}(d_{ij} - \lambda d_{ij}) \) for a point in the Swiss roll centre:

\( \lambda = 10^{-4} \): Gaussian

\( \lambda = 1 \): Mexican hat

Zoom view:

3 The Elastic Embedding (EE)

We can achieve a similar point-separating effect with a simpler prior. Besides, we make it data-dependent and control its strength with a parameter \( \lambda \in \mathbb{R} \):

\[
\min \sum_{i,j} w_{ij} (\|x_i - x_j\|^2) + \lambda \sum_{i,j} w_{ij} \exp(-\|x_i - x_j\|^2)
\]

where \( w_{ij} = \Omega_{ij} - y_i - y_j \) and we have two graphs:

\begin{itemize}
  \item One with attractive weights \( W' = \Omega_{ij} \) (Gaussian affinities, geodesic distances, etc.).
  \item One with repulsive weights \( W'' = \Omega_{ij} \) (i.e. \( \Omega_{ij} \) so \( w_{ij} = \Omega_{ij} - y_i - y_j \)).
\end{itemize}

EE symmetries the constraints of Laplacian eigenmaps, where both types of mistakes are penalised: placing for apart latent points that correspond to similar data points, and placing close together latent points that correspond to dissimilar data points.

Minimising \( E \):

Several options:

1. Homotopy: increase \( \lambda \) from \( \lambda^* \) sufficiently slowly while minimising \( E \) over \( X \); achieves very good optima, but is slow.

2. Fixed- \( \lambda \): take \( \lambda \) from a spectral method; fast, but optimum dependent on initial \( X \).

3. Homotopy with \( \lambda \) s.t. quadratic constraints on \( X \) as in LE: \( X = \lambda L^{-1} \) for \( \lambda > 0 \), which is a better initial point than \( X = 0 \); but more difficult constrained optimisation.

Efficient optimisation (faster than SNE) by using search directions derived from a fixed-point equation (see paper). Cost per iteration: \( O(L\lambda^2) \) (or \( O(L\lambda^2) \) with sparse graphs).

Out-of-sample mappings

Given a new point \( x \in \mathbb{R}^D \), we solve the original EE problem over \( X(x) \) and \( Y(y) \) s.t. keeping the embedding \( X \) fixed:

\[
E(X(x), Y(y)) = \sum_{i,j} w_{ij} (\|y_i - x \|^2 - \lambda \|y_i - x\|^2) + \lambda \sum_{i,j} w_{ij} \exp(-\|y_i - x\|^2)
\]

with kernels induced from the affinity kernels that were used in the EE training (using the same neighbourhood structure):

\[
w'_{ij}(y) = \exp(-\|y_i - y_j\|^2) \quad w''_{ij}(y) = \exp(-\|y_i - y_j\|^2)
\]

Project: \( F(y) = \arg \min_{F} E(X,f(Y,y)) \). Reconstruct: \( X = \arg \min_{X} E(X,f(Y,y)) \). Nonparametric (implicit) solution with the form of a nonconvex t.c. of \( (Y,X) \) (the weights can be negative):

\[
F(y) = x - \sum_{i,j} w_{ij}(y) \tilde{x}_i
\]

This allows to extrapolate beyond the dataset.

4 Behaviour of the embedding \( X(\lambda) \) as a function of \( \lambda \)

- Write \( X(\lambda) = (x_1, \ldots, x_N) \). The embedding \( X(\lambda) \) satisfies the stationary point equation

\[
\frac{\partial}{\partial x} E_X = \mathbf{0} = (X - \lambda X L^{-1} X) - \sum_{i,j} w_{ij} (y_i - y_j) \exp(-\|y_i - y_j\|^2)
\]

\( W \) are learned affinities and can be negative if \( \lambda \) is large enough.

\( X \) is in the nullspace of the learned graph Laplacian \( L - \lambda I \) (very similar to a 1D LE embedding).

- The repulsion in the prior term in \( E \) drops off rapidly, so the embedding \( X \) has a characteristic scale at each \( \lambda \) that increases with \( \lambda \).

- The embedding \( X(\lambda) \) undergoes a series of bifurcations where an eigenvalue of the Hessian of \( E \) becomes negative. At some of these \( \lambda \), \( X \) increases.

- As \( \lambda \) increases, \( X \) unfolds globally and reorganises locally, representing better the global and local manifold structure. This is the region of \( \lambda \) where the best embeddings occur (region 0 in the Swiss roll).

- For large \( \lambda \), the points distribute approximately equidistant from each other locally (first grid in 2D) while maintaining the global structure of the manifold, and the map scale increases logarithmically with \( \lambda \).

- The learned affinities start as Gaussian for small \( \lambda \) and become Mexican-hat functions as \( \lambda \) increases.

We illustrate this for a Swiss roll dataset with \( N = 2000 \) points in \( D = 3 \).

5 Experiments: COIL–20 dataset

Rotation sequences of \( N \) objects every \( \psi \) degrees; each data point is a greyscale image of \( 128 \times 128 \), so \( Y \) has \( N = 720 \) points in \( D = 16 \times 24 \) dimensions. All methods randomly initialised. For EE: \( \lambda = 1 \); out-of-sample result obtained by training on half the data (even-numbered images) and testing on the other half.

6 Conclusion

Our motivation is that nonlinear manifold learning algorithms can far outperform spectral methods if we design simple, meaningful objective functions and we find good local optima efficiently. We demonstrate this with the elastic embedding (EE) and show that it learns at the same time the embedding and, implicitly, the affinities; the latter are responsible for the SNE's successes and might perform better than Gaussian affinities with spectral methods. Many of these insights (homotopy, search directions, out-of-sample mappings, learned affinities) carry over to SNE and related methods.

Supported by NSF CAREER award IIS–0754089.