The Elastic Embedding algorithm for dimensionality reduction



Miguel Á. Carreira-Perpiñán Electrical Engineering and Computer Science University of California, Merced http://eecs.ucmerced.edu

Laplacian eigenmaps (LE) (Belkin & Niyogi 2002)

Given affinities W (e.g. Gaussian) for data points $y_1, \ldots, y_N \in \mathbb{R}^D$, their latent coordinates $X_{L \times N} = (x_1, \ldots, x_N)$ are obtained as:

 $\min E_{\mathsf{LE}}(\mathbf{X}) = \sum_{n,m=1}^{N} w_{nm} \|\mathbf{x}_n - \mathbf{x}_m\|^2 \text{ s.t. translation \& scale constraints}$

- Discourages placing far apart latent points x_n , x_m that correspond to similar data points y_n , y_m , but places no direct constraint on pairs associated with distant data points.
- While it can capture the global structure of the manifold, it often leads to distorted maps, particularly if multiple manifolds exist: large clusters of points collapse, local clusters and gaps, boundary effects.

Solution Global optimum from spectral problem: eigenvectors of the $N \times N$ graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$ with $\mathbf{D} = \text{diag}(\sum_{m} w_{nm})$.

Laplacian eigenmaps: Swiss roll



X from Laplacian eigenmaps



Laplacian eigenmaps: COIL-20 dataset

Rotations of 20 objects every 5°; $\mathbf{y}_n = \text{greyscale image of } 128 \times 128$.

X from Laplacian eigenmaps

X from SNE



Plots from van der Maaten & Hinton 2008

Stochastic neighbour embedding (SNE) (Hinton & Roweis 2003)

$$p_{nm} = \frac{\exp\left(-d_{nm}^{2}\right)}{\sum_{n \neq m'} \exp\left(-d_{nm'}^{2}\right)}, \ p_{nn} = 0; \quad q_{nm} = \frac{\exp\left(-\|\mathbf{x}_{n} - \mathbf{x}_{m}\|^{2}\right)}{\sum_{n \neq m'} \exp\left(-\|\mathbf{x}_{n} - \mathbf{x}_{m'}\|^{2}\right)}$$
$$E_{\mathsf{SNE}}(\mathbf{X}) = \sum_{n=1}^{N} D\left(P_{n}\|Q_{n}\right) = \sum_{n,m=1}^{N} p_{nm} \log \frac{p_{nm}}{q_{nm}}$$

- Tries to match the latent-space distributions Q_n over neighbours to the data-space ones P_n .
- Significantly better embeddings than LE, particularly when multiple manifolds exist.

E is very nonlinear: local optima, difficult optimisation.
 Hinton & Roweis 2003: "[The SNE] cost function cleanly enforces both keeping the images of nearby objects nearby and keeping the images of widely separated objects relatively far apart".

A relation between LE and SNE

But in fact $E_{SNE}(\mathbf{X})$ equals (up to constants):



♦ Term ① is like LE.

Term ② is a "prior" that pushes apart all latent point pairs equally, irrespectively of whether their high-dimensional counterparts are close or far in data space.

So SNE enforces keeping the images of nearby objects nearby (like LE) while pushing all images apart from each other. This prior is what makes SNE improve significantly over LE.

A better point-separating prior term

The previous observation suggests a better "prior":

- a simpler expression with a similar point-separating effect
- \diamond data-dependent: separate $\mathbf{x}_n, \mathbf{x}_m$ for distant $\mathbf{y}_n, \mathbf{y}_m$

* strength controlled with a parameter $\lambda \ge 0$

which we define as follows:

$$\lambda \sum_{n,m=1}^{N} \overline{w}_{nm}^{-} \|\mathbf{y}_{n} - \mathbf{y}_{m}\|^{2} \exp\left(-\|\mathbf{x}_{n} - \mathbf{x}_{m}\|^{2}\right)$$

where \overline{w}_{nm}^- are graph weights, possibly sparse.

The Elastic Embedding (EE) model

$$\min E(\mathbf{X}; \lambda) = \sum_{n,m=1}^{N} w_{nm}^{+} \|\mathbf{x}_{n} - \mathbf{x}_{m}\|^{2} + \lambda \sum_{n,m=1}^{N} w_{nm}^{-} \exp\left(-\|\mathbf{x}_{n} - \mathbf{x}_{m}\|^{2}\right)$$

where $w_{nm}^- = \overline{w}_{nm}^- \|\mathbf{y}_n - \mathbf{y}_m\|^2$ and we have two graphs:

- One with attractive weights $W^+ = (w_{nm}^+)$ (normalised) Gaussian affinities, geodesic distances, etc.
- One with repulsive weights $\mathbf{W}^- = (w_{nm}^-)$ e.g. $\overline{w}_{nm}^- = 1$ so $w_{nm}^- = \|\mathbf{y}_n - \mathbf{y}_m\|^2$.

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



p. <u>8</u>



EE,
$$\lambda = 10^1$$





LE

SNE



 $\text{True } \mathbf{X}$



Learned affinities $w_{nm} = w_{nm}^+ - \lambda \widetilde{w}_{nm}^-$ for a point in the Swiss roll centre: $\lambda = 10^{-2}$: Gaussian $\lambda = 10^1$: Mexican-hat



Zoom view:



Zoom view:





* The embedding $\mathbf{X}(\lambda)$ satisfies the stationary point equation

$$\frac{\partial E}{\partial \mathbf{X}} = 4\mathbf{X}\mathbf{L} = 4\mathbf{X}(\underbrace{\mathbf{D}^{+} - \mathbf{W}^{+}}_{\mathbf{L}^{+}} - \lambda(\underbrace{\widetilde{\mathbf{D}^{-} - \widetilde{\mathbf{W}^{-}}}_{\mathbf{L}^{-}})) = \mathbf{0}$$

$$\underbrace{\mathbf{U}_{nm}}_{\mathbf{M}} = w_{nm}^{+} - \lambda\widetilde{w}_{nm}^{-} \qquad \widetilde{w}_{nm}^{-} = w_{nm}^{-}\exp\left(-\|\mathbf{x}_{n} - \mathbf{x}_{m}\|^{2}\right)$$

- ♦ W are learned affinities, and define a learned graph Laplacian L = L⁺ - $\lambda \widetilde{L}^-$.
- If λ is large enough, W is not positive definite and has some negative entries.
- \bullet X is in the nullspace of L ("spectral problem").
- * The repulsion in the prior term in E drops off rapidly, so the embedding X has a characteristic scale at each λ that increases with λ . The initial X should have the right scale.

- $\mathbf{X}(\lambda)$ undergoes a series of bifurcations where an eigenvalue of the Hessian of *E* becomes negative. At some of these dim (\mathbf{X}) increases.
- At the first bifurcation λ = λ₁^{*}, dim (**X**) grows from 0 to 1, and **X** expands along the trailing eigenvector of L⁺ − λ₁^{*}L[−] (similar to a 1D LE embedding). We have upper and lower bounds for λ₁^{*}:

$$\max\left(\frac{\lambda_2^+}{\lambda_N^-}, \min_{n,m} \frac{w_{nm}^+}{w_{nm}^-}\right) \le \lambda_1^* \le \min\left(\frac{\lambda_2^+}{\lambda_2^-}, \dots, \frac{\lambda_N^+}{\lambda_N^-}, \frac{L_{11}^+}{L_{11}^-}, \dots, \frac{L_{NN}^+}{L_{NN}^-}\right).$$

- As λ increases, X unfolds globally and reorganises locally, representing better the global and local manifold structure. This is the region of λ where the best embeddings occur.
- For large λ, the points distribute approximately equidistant from each other locally (hex grid in 2D) while maintaining the global structure of the manifold, and the map scale increases logarithmically with λ.
- The learned affinities start as Gaussian for small λ and become Mexican-hat functions as λ increases.

Several options:

- 1. Homotopy:
 - * increase λ from λ_1^* suffic. slowly while minimising *E* over **X**
 - achieves very good optima, but is slow.
- 2. Fixed λ :
 - init. X from a spectral method (rescaled to match the scale at λ)
 - \diamond fast, but optimum dependent on initial X.
- 3. Homotopy with E s.t. quadratic constraints on X as in LE:
 - * X = LE for $\lambda = 0$, which is a better initial point than X = 0
 - more difficult constrained optimisation.

Other than not increasing λ too fast, no special user parameters required (momentum or learning rate, amount of jitter, etc.)

Gradient descent, conjugate gradients: very slow, requires tiny steps (this also applies to SNE); reason: *E* is ill-conditioned. Search directions derived from a fixed-point equation (a la Jacobi, Gauss-Seidel, etc.):

- ♦ Fixed-point it. X ← -XBA⁻¹ not convergent ⇒ use direction $\Delta = -XBA^{-1} X \text{ in a line search } X \leftarrow X + η\Delta \text{ for } η \ge 0.$
- ♦ Δ is descent. If cond (A) upper bounded + I.s. satisfies Wolfe conditions ⇒ global convergence (from any initial $X_0 \in \mathbb{R}^{L \times N}$).
- Some splittings are many times faster than gradient descent. In particular $A = 4D^+$ (computable in $\mathcal{O}(NL)$) requires no line search at all ($\eta = 1$) except when λ is close to a bifurcation.

Ongoing: even better directions with a diag. pd approx. to Hessian. Cost per iteration: $\mathcal{O}(LN^2)$, or $\mathcal{O}(LN)$ with sparse graphs.

Out-of-sample mappings

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

$$E'(\mathbf{x}, \mathbf{y}) = 2\sum_{n=1}^{N} \left(w^+(\mathbf{y}, \mathbf{y}_n) \|\mathbf{x} - \mathbf{x}_n\|^2 + \lambda w^-(\mathbf{y}, \mathbf{y}_n) \exp\left(-\|\mathbf{x} - \mathbf{x}_n\|^2\right) \right)$$

init. to the closest (x_n, y_n) , with kernels induced from the affinity kernels that were used in the EE training (using the same neighbourhood structure):

$$w^+(\mathbf{y},\mathbf{y}_n) = \exp\left(-\frac{1}{2} \left\| (\mathbf{y} - \mathbf{y}_n) / \sigma \right\|^2 \right) \qquad w^-(\mathbf{y},\mathbf{y}_n) = \widetilde{w}_n^- \left\| \mathbf{y} - \mathbf{y}_n \right\|^2$$

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

$$\mathbf{F}(\mathbf{y}) = \mathbf{x} = \sum_{n=1}^{N} \frac{w_n(\mathbf{x})}{\sum_{n'=1}^{N} w_{n'}(\mathbf{x})} \mathbf{x}_n \qquad \mathbf{f}(\mathbf{x}) = \mathbf{y} = \sum_{n=1}^{N} \frac{\zeta_n(\mathbf{y})}{\sum_{n'=1}^{N} \zeta_{n'}(\mathbf{y})} \mathbf{y}_n.$$

This allows to extrapolate beyond the dataset.

Optimisation: Gauss-Newton for F, search directions from fixed-point iteration or diagonal pd Hessian approx. for f. Beats gradient descent.

Example

COIL-20 dataset

Y: $(N = 720) \times (D = 16384)$. All methods randomly initialised. L = 2.

X from EE ($\lambda = 1$)



 ${\bf X}$ from SNE



 \mathbf{X} from t-SNE



COIL-20 dataset: **EE** out-of-sample mapping results

Training: even-numbered images; testing: odd-numbered images.

$$\mathbf{F}: \mathbf{y} = \mathbf{E} \in \mathbb{R}^{16\,384} \to \mathbf{x} \in \mathbb{R}^2$$





Conclusions

- Our motivation: nonlinear manifold learning algorithms can far outperform spectral methods if we design simple, meaningful objective functions and we find good local optima efficiently.
- We show this with a new method, the elastic embedding. EE deals symmetrically with the data and latent points, penalising placing far apart latent points from similar data points, and placing close together latent points from dissimilar data points.
- EE learns at the same time the embedding and, implicitly, the affinities; the latter resemble Mexican-hat functions and might perform better than Gaussian affinities with spectral methods.
- Our insights carry over to SNE, *t*-SNE and related methods: homotopy over λ, efficient search directions with global convergence, out-of-sample mappings, learned affinities.

Matlab code: http://eecs.ucmerced.edu Work supported by NSF CAREER award IIS-0754089.