The Elastic Embedding Algorithm for Dimensionality Reduction

Miguel Á. Carreira-Perpiñán MCARREIRA-PERPINAN@UCMERCED.EDU
Electrical Engineering and Computer Science, School of Engineering, University of California, Merced

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 betwen 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, characterise 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 those of SNE, but more efficiently and robustly.

1. Introduction

We consider the problem of dimensionality reduction, also called manifold learning, where we seek to explain an observed high-dimensional data set $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$ of $D \times N$ in terms of a much smaller number of dimensions $L \ll D$. We will focus on algorithms that take as input $\mathbf{Y}$ and estimate only the coordinates of the corresponding low-dimensional (latent) points $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ of $L \times N$. Within this framework, there has been an enormous amount of recent work on spectral dimensionality reduction, mainly algorithmic but also theoretical. Spectral methods such as Isomap (Tenenbaum et al., 2000), LLE (Roweis & Saul, 2000) or Laplacian eigenmaps (Belkin & Niyogi, 2003) formulate an objective function of $\mathbf{X}$ based on pairwise affinities defined on a neighbourhood graph of the data, whose solution is given by the nontrivial extremal eigenvectors of a certain $N \times N$ matrix, often sparse. Their success is due to their lack of local optima, to the availability of excellent numerical eigensolvers, and to the remarkably good results (often capturing the global structure of the underlying manifold) that can be obtained with simple affinity functions such as Gaussian. That said, such simple affinities can only do so much, and the maps obtained typically collapse non-similar points in small latent space regions or, conversely, leave large gaps in it; they also show significant boundary effects. These problems are particularly clear when the data consists of several separate manifolds. Using more sophisticated affinities that encode non-local information (as in Isomap or in maximum variance unfolding; Weinberger & Saul, 2006) can improve this at a higher computational cost, but are still sensitive to noise in the data and in the graph.

Nonlinear methods such as stochastic neighbour embedding (SNE) can go beyond spectral methods and find better optima, representing the global and local structure as well as dealing with multiple manifolds (Hinton & Roweis, 2003; van der Maaten & Hinton, 2008). However, nonlinear methods are far less developed. Only a few such methods have been proposed, and while their results are very encouraging, their optimisation is costly and prone to local optima, and our understanding of the algorithms is limited to an intuitive interpretation of their objective function. The objective of this paper is (1) to propose a new nonlinear algorithm of this type, with results as good as those of SNE but more efficient and robust; and (2) to further our understanding of this type of algorithms.

2. Related work

Metric MDS (Borg & Groenen, 2005) preserves data-space distances in latent space by minimising an objective function (stress) of the latent coordinates. The linear version (classical scaling) results in an eigenvalue problem with a unique solution in generic cases. Several nonlinear versions ex-

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

ist, such as the Sammon mapping (Sammon, 1969), which are generally difficult to optimise and prone to bad local optima. Stochastic neighbour embedding (SNE) (Hinton & Roweis, 2003) preserves probabilities instead of distances, and earlier papers (Hinton & Roweis, 2003; van der Maaten & Hinton, 2008) have shown its superiority over other MDS-like methods when dealing with data that lies in nonlinear, clustered manifolds (though the optimisation is still difficult). We will focus on SNE-type methods. SNE defines the following normalised, non-symmetric affinities p_{nm} and q_{nm} for each data point n in the data and latent spaces, respectively:

$$p_{nm} = \frac{\exp\left(-d_{nm}^2\right)}{\sum_{n\neq m'} \exp\left(-d_{nm'}^2\right)} \qquad p_{nn} = 0 \tag{1}$$

$$q_{nm} = \frac{\exp\left(-\left\|\mathbf{x}_n - \mathbf{x}_m\right\|^2\right)}{\sum_{n\neq m'} \exp\left(-\left\|\mathbf{x}_n - \mathbf{x}_{m'}\right\|^2\right)}. \tag{2}$$

We will take $d_{nm}^2 = \frac{1}{2}\left\|(\mathbf{y}_n - \mathbf{y}_m)/\sigma_n\right\|^2$, that is, Gaussian affinities, though other types of affinity may be used. Each width σ_n is chosen by a binary search so the entropy of the distribution P_n over neighbours is roughly $\log k$ (for a user-provided $k \ll N$, which is then the perplexity, or effective number of neighbours). SNE minimises the following objective function:

$$E_{\mathrm{SNE}}(\mathbf{X}) = \sum_{n=1}^{N} \mathrm{D}\left(P_n \| Q_n\right) = \sum_{n,m=1}^{N} p_{nm} \log \frac{p_{nm}}{q_{nm}} \tag{3}$$

and so tries to match the latent-space distributions over neighbours to the data-space ones.

One important disadvantage of SNE is that its gradient-based optimisation is slow and requires care to find good optima. The user must tune for each dataset several parameters (learning rate, momentum rate, amount of gradient jitter, etc., and all these must be adapted by hand as the optimisation proceeds). Some versions of SNE have been proposed that slightly simplify the gradient by symmetrising the objective function (Venna & Kaski, 2007) or the probabilities (Cook et al., 2007), but the normalisation term in the q_{nm} terms still makes it very nonlinear.

When the dimensionality L of the latent space is smaller than the intrinsic dimensionality of the data, the resulting map is unavoidably distorted. For purposes of visualisation (rather than of faithful dimensionality reduction), Cook et al. (2007) and van der Maaten & Hinton (2008) have proposed two ways of improving the separation of clusters in this case, UNI-SNE and t-SNE, resp. UNI-SNE biases each q_{nm} by a fixed constant, while t-SNE defines Q as a t-distribution with one degree of freedom. In both cases Q has longer tails, which allows $\mathbf{x}$-points corresponding to $\mathbf{y}$-pairs at a moderate distance to separate more.

SNE and our EE (described later) are related with the elastic net (EN) (Durbin et al., 1989), originally proposed to approximate the travelling salesman problem. The EN minimises the sum of a log-sum term (Gaussian-mixture likelihood) that moves $\mathbf{Y}$-space centroids towards the data, and a data-independent quadratic prior on the centroids (a graph Laplacian prior), using a homotopy method. In the EN, the centroids move in the data, rather than the latent, space; and the quadratic prior on them enforces a predetermined topology rather than being based on data affinities. An interesting connection with EE (see section 6), is that the EN prior is provably equivalent to a certain Mexican-hat interaction term (Carreira-Perpiñán & Goodhill, 2004).

EE (and SNE) can be seen as symmetrising 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. A related phenomenon occurs with principal curves (Hastie & Stuetzle, 1989) and Dimensionality Reduction by Unsupervised Regression (Carreira-Perpiñán & Lu, 2008); the latter may be seen as a symmetrised version of the former that penalises errors in the data *and* the latent space. Closing the loop in this way seems to lead to better embeddings.

In the rest of the paper, we show a relation between SNE and Laplacian eigenmaps (sec. 3) that immediately suggests our EE algorithm, which we then study (sec. 4) and apply in practice (sec. 5).

3. A Relation between SNE and LE

Laplacian eigenmaps (LE) (Belkin & Niyogi, 2003) is a spectral method that optimises

$$E_{\mathrm{LE}}(\mathbf{X}) = \sum_{n,m=1}^{N} w_{nm} \left\|\mathbf{x}_n - \mathbf{x}_m\right\|^2 \tag{4}$$

subject to quadratic and linear constraints, and has a unique solution given by the nontrivial trailing eigenvectors of the normalised version of the graph Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{W}_{N\times N}$ is the symmetric affinity matrix (typically Gaussian) and $\mathbf{D} = \mathrm{diag}\left(\sum_{n=1}^{N} w_{nm}\right)$ the degree matrix. LE discourages placing far apart latent points that correspond to similar data points, but places no direct constraint on pairs associated with distant data points. This often leads to distorted maps where large clusters of points col-

lapse (as happens with related methods such as LLE).

There is a fundamental relation between LE and SNE. Expanding (3) and ignoring terms that do not depend on $\mathbf{X}$, we have that

$$E_{\text{SNE}}(\mathbf{X}) = \sum_{n,m=1}^{N} p_{nm} \|\mathbf{x}_n - \mathbf{x}_m\|^2 + \sum_{n=1}^{N} \log \sum_{n \neq m} \exp\left(-\|\mathbf{x}_n - \mathbf{x}_m\|^2\right) \quad (5)$$

since $\sum_{m=1}^{N} p_{nm} = 1$ for each n. The first term in the RHS is identical to the LE objective if using normalised affinities as in diffusion maps (i.e., taking $w_{nm} = p_{nm}$). It is a *local distance term*, and also a *data-dependent term* (since it depends on the data $\mathbf{Y}$ through the p_{nm}). The second term encourages latent points to separate from each other as much as possible (or until the exponentials become negligible). It is a *global distance term*, symmetric wrt $\mathbf{x}_n$ and $\mathbf{x}_m$, which pushes apart all point pairs equally, irrespective of whether their high-dimensional counterparts are close or far in data space. It is also a *data-independent term*, since it does not depend on the data $\mathbf{Y}$.

Therefore, SNE may be seen as LE with a data-independent prior that blows points apart from each other. It is thus more accurate to say that the SNE objective function enforces keeping the images of nearby objects nearby while pushing all images apart from each other, rather than to say that it enforces both keeping the images of nearby objects nearby and keeping the images of widely separated objects relatively far apart. However, this prior does cause the result from SNE to be radically different from that of LE, improving the spacing of points and clusters, and better representing the manifold structure. We are now ready to introduce our algorithm.

4. The Elastic Embedding (EE)

We define the objective function

$$E(\mathbf{X}; \lambda) = \sum_{n,m=1}^{N} w^{+}_{nm} \|\mathbf{x}_n - \mathbf{x}_m\|^2 + \lambda \sum_{n,m=1}^{N} w^{-}_{nm} \exp\left(-\|\mathbf{x}_n - \mathbf{x}_m\|^2\right) \quad (6)$$

where $w^{-}_{nm} = \overline{w}^{-}_{nm} \|\mathbf{y}_n - \mathbf{y}_m\|^2$ and we have two graphs: one with *attractive* weights $\mathbf{W}^{+} = (w^{+}_{nm})$ and the other with *repulsive* weights $\mathbf{W}^{-} = (w^{-}_{nm})$, both nonnegative. The left (+) term is the LE term and preserves local distances, where w^{+}_{nm} could be (normalised) Gaussian affinities, geodesic distances, commuting times or other affinities, possibly nonsymmetric or sparse. The right (−) term preserves global distances or separates latent points as in SNE *but in a simpler way*. This repulsion becomes negligible once neigbouring $\mathbf{x}$s are farther apart than a characteristic, λ-dependent scale, so the map remains somewhat compact. The regularisation parameter $\lambda \geq 0$ trades off both terms. For simplicity, consider full graphs $w^{+}_{nm} = \exp\left(-\frac{1}{2}\|(\mathbf{y}_n - \mathbf{y}_m)/\sigma\|^2\right)$ and $\overline{w}^{-}_{nm} = 1$ $\forall n \neq m$, with $w^{+}_{nn} = w^{-}_{nn} = 0$ $\forall n$; although some of our results use sparse graphs. Note that the $\mathbf{X}$ resulting from EE are equivalent up to rigid motions, and that globally rescaling the data simply rescales λ: $E(\mathbf{X}; \lambda; \mathbf{Y}, \sigma) = E(\mathbf{X}; \lambda/\alpha^2; \alpha\mathbf{Y}, \alpha\sigma)$.

We can then obtain the gradient of E from eq. (6):

$$\frac{\partial E}{\partial \mathbf{x}_n} = 4 \sum_{m \neq n}^{N} w_{nm}(\mathbf{x}_n - \mathbf{x}_m) \quad (7)$$

$$\mathbf{G}(\mathbf{X}; \lambda) = \frac{\partial E}{\partial \mathbf{X}} = 4\mathbf{X}(\mathbf{L}^{+} - \lambda \widetilde{\mathbf{L}}^{-}) = 4\mathbf{X}\mathbf{L} \quad (8)$$

where we define the affinities

$$\widetilde{w}^{-}_{nm} = w^{-}_{nm} \exp\left(-\|\mathbf{x}_n - \mathbf{x}_m\|^2\right) \quad (9)$$

$$w_{nm} = w^{+}_{nm} - \lambda \widetilde{w}^{-}_{nm} \quad (10)$$

and their graph Laplacians $\widetilde{\mathbf{L}} = \widetilde{\mathbf{D}} - \widetilde{\mathbf{W}}$, $\mathbf{L} = \mathbf{D} - \mathbf{W}$ in the usual way. Note that $\mathbf{L}^{+}$ is the usual (unnormalised) graph Laplacian that appears in Laplacian eigenmaps. $\mathbf{W}$ can be considered a *learned affinity* matrix and contains negative weights for $\lambda > 0$. Both the objective function and the gradient of EE are quite less nonlinear than those of SNE and its variations because we have eliminated the cumbersome log-sum term. This results in an easier optimisation and presumably fewer local optima.

At a minimiser (for each λ) we have $\mathbf{G}(\mathbf{X}; \lambda) = \mathbf{0}$, so the embedding $\mathbf{X}(\lambda)$ satisfies $\mathbf{X}\mathbf{L} = \mathbf{0}$ and therefore consists of eigenvectors of the nullspace of the graph Laplacian $\mathbf{L}$ for the *learned* graph affinity matrix $\mathbf{W}$. In minimising E at each λ, we both construct this graph and find its nullspace eigenvectors (a spectral problem). Note that this does not mean that EE at a given λ is equivalent to LE using as affinity matrix $\mathbf{W}$, as LE would find the eigenvectors associated with the algebraically smallest eigenvalues, which for large enough λ are negative.

4.1. Study of the case $N = 2$

The simple case of $N = 2$ points in 1D is surprisingly informative. Take w.l.o.g. one point at the origin and

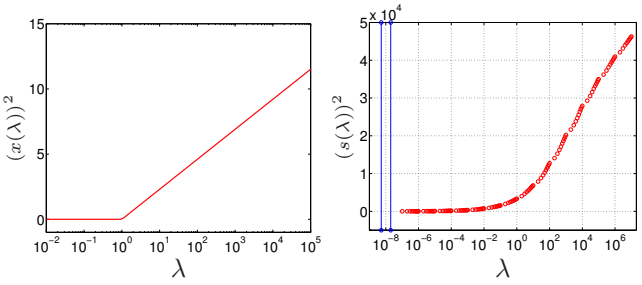

Figure 1. Left: plot of $(x(\lambda))^2$ when the dataset has $N = 2$ points in 1D, for $w^+ = w^- = \lambda_1^* = 1$. *Right*: plot of the squared diameter of $\mathbf{X}(\lambda)$ for the Swiss roll example (blue vertical lines bound λ_1^*).

call $x \geq 0$ the position of the other. The objective function for $\lambda \geq 0$ is $E(x;\lambda) = 2(w^+x^2 + \lambda w^- e^{-x^2})$. Define the critical value $\lambda_1^* = w^+/w^-$ at which a bifurcation occurs. For $\lambda < \lambda_1^*$, E has a minimum at $x = 0$ and nothing interesting happens, the points are coincident. For $\lambda > \lambda_1^*$ it has a maximum at $x = 0$ and a minimum at $x = \sqrt{\log(\lambda/\lambda_1^*)}$; as λ grows, the points separate very slowly (square-root-log). Fig. 1 (left) shows the squared scale of the map $(x(\lambda))^2 = \max(0, \log(\lambda/\lambda_1^*))$, which grows logarithmically after the bifurcation. When $N > 2$, $\mathbf{X}(\lambda)$ follows the same behaviour for small or large λ (fig. 1 right), but shows multiple bifurcations and local minima for intermediate λ.

4.2. Study of the critical $\lambda = \lambda_1^*$ for $N > 2$

For $\lambda \leq \lambda_1^*$, the EE objective function $E(\mathbf{X};\lambda)$ is minimised by the point $\mathbf{X} = \mathbf{0}$, corresponding to an embedding where all latent points are coincident. For $\lambda > \lambda_1^*$, the embedding unfolds. We want to locate this bifurcation and study the evolution of the point $\mathbf{X} = \mathbf{0}$ at it. $\mathbf{X} = \mathbf{0}$ is a stationary point for all λ, and the Hessian of $E(\mathbf{X};\lambda)$ at it equals

$$\mathbf{H}(\mathbf{0};\lambda) = 4\,\mathrm{diag}\left(\mathbf{L}^+ - \lambda\mathbf{L}^-, \dots, \mathbf{L}^+ - \lambda\mathbf{L}^-\right)$$

since $\widetilde{\mathbf{L}}^- = \mathbf{L}^-$, and assuming $\mathbf{X}$ in row-major order. Using Taylor's theorem, we can approximate $E(\mathbf{X};\lambda)$ near $\mathbf{X} = \mathbf{0}$ to second order. Since the Hessian is block-diagonal with $N \times N$ blocks corresponding to the L dimensions of $\mathbf{X}$, we can study the behaviour of each dimension separately by studying the function $e(\boldsymbol{x};\lambda) = \boldsymbol{x}^T\mathbf{L}\boldsymbol{x}$, where $\mathbf{L} = \mathbf{L}^+ - \lambda\mathbf{L}^-$ and $\boldsymbol{x}$ is an $N \times 1$ vector containing the coordinates of all N points in dimension 1 (i.e., the first row of $\mathbf{X}$ transposed). The function e is the sum of two quadratic functions of opposite sign and the negative one is weighted by λ. For $\lambda = 0$, e equals the positive quadratic, for $\lambda \to \infty$ it tends to the negative one, and for intermediate λ, e is a hyperbolic paraboloid. For $\lambda \leq \lambda_1^*$, e is positive semidefinite and has global minima along the line $\boldsymbol{x} = \alpha\mathbf{u}_0$ for $\alpha \in \mathbb{R}$ and $\mathbf{u}_0 = \mathbf{1}$; this is the eigenvector of $\mathbf{L}$ associated with a null eigenvalue, and represents the embedding where all points are coincident; we will refer to this embedding as $\boldsymbol{x} = \mathbf{0}$ for simplicity. Eigenvector $\mathbf{1}$ exists for all λ and represents EE's invariance to global translations of $\mathbf{X}$. We are interested in the critical value λ_1^* when $\mathbf{L}$ stops being positive semidefinite and the hyperbolic paraboloid first arises. At that point, $\boldsymbol{x} = \mathbf{0}$ is about to stop being a minimum and become a saddle point (or a maximum in particular cases), and a second eigenvector $\mathbf{u}_1^*$ exists with null eigenvalue and orthogonal to $\mathbf{u}_0$. For $\lambda > \lambda_1^*$, $\mathbf{u}_1^*$ is associated with a negative eigenvalue, and e decreases fastest along $\mathbf{u}_1^*$. Thus, when λ is just larger than λ_1^*, each dimension of the embedding expands along the negative eigenvector $\mathbf{u}_1^*$ of $\mathbf{L}$. Note that $\mathbf{u}_1^*$ is the second trailing eigenvector of $\mathbf{L}$ at $\lambda = \lambda_1^*$ and thus corresponds to the 1D embedding that LE would produce with an affinity matrix $\mathbf{W}^+ - \lambda_1^*\mathbf{W}^-$ and constraints $\mathbf{X}\mathbf{X}^T = \mathbf{I}$ and $\mathbf{X}\mathbf{1} = \mathbf{0}$; in practice this is close to the first nontrivial trailing eigenvector of $\mathbf{L}^+$.

In summary, at the bifurcation $\lambda = \lambda_1^*$, the latent points separate and an embedding $\mathbf{X}$ arises that is 1D and very similar to the 1D LE embedding. This embedding $\mathbf{X}(\lambda)$ keeps evolving as λ is increased.

We do not have an explicit form for λ_1^* or $\mathbf{u}_1^*$, but we have upper and lower bounds $l_1 \leq \lambda_1^* \leq u_1$ that are quite tight in practice (proof omitted for lack of space):

$$l_1 = \max\left(\frac{\lambda_2^+}{\lambda_N^-}, \min_{n,m}\frac{w_{nm}^+}{w_{nm}^-}\right) \tag{11}$$

$$u_1 = \min\left(\frac{\lambda_2^+}{\lambda_2^-}, \dots, \frac{\lambda_N^+}{\lambda_N^-}, \frac{L_{11}^+}{L_{11}^-}, \dots, \frac{L_{NN}^+}{L_{NN}^-}\right). \tag{12}$$

4.3. Minimising $E(\mathbf{X};\lambda)$ for fixed λ

We have noticed that minimising E with gradient descent or conjugate gradients is very slow and requires tiny steps (this also applies to SNE). Using search directions derived from a fixed-point iteration works much better. Rearranging the stationary point equation (a matrix of $L \times N$)

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

as a splitting $\mathbf{G} = \mathbf{X}(\mathbf{A} + \mathbf{B}) = \mathbf{0}$, where $\mathbf{A}$ is symmetric positive definite and $\mathbf{B}$ is symmetric, we obtain a fixed point iteration $\mathbf{X} \leftarrow -\mathbf{X}\mathbf{B}\mathbf{A}^{-1}$. Although this iteration does not always converge, it does suggest using a search direction $\boldsymbol{\Delta} = -\mathbf{X}\mathbf{B}\mathbf{A}^{-1} - \mathbf{X} = -\mathbf{X}(\mathbf{B}\mathbf{A}^{-1} + \mathbf{I}) = -\mathbf{G}\mathbf{A}^{-1}$ along which we can decrease E with a line search $\mathbf{X} \leftarrow \mathbf{X} + \eta\boldsymbol{\Delta}$ for $\eta \geq 0$.

The direction $\mathbf{\Delta}$ is descent and, if $\operatorname{cond}(\mathbf{A})$ is upper bounded, it never becomes too close to being orthogonal to the gradient (proof omitted). If the line search satisfies e.g. the Wolfe conditions, then, by Zoutendijk's theorem (th. 3.2 in Nocedal & Wright, 2006), the algorithm converges to a stationary point of E from any initial $\mathbf{X}_0 \in \mathbb{R}^{L\times N}$. We have tried several splittings and found them to improve greatly over the gradient ($\mathbf{A}=\mathbf{I}$), in particular $\mathbf{A}=4\mathbf{D}^+$ (computable in $\mathcal{O}(NL)$). In practice, we find this requires no line search at all ($\eta=1$) except when λ is close to a bifurcation, most notably near λ_1^*.

The cost of each iteration is thus $\mathcal{O}(LN^2)$ (or $\mathcal{O}(LN)$ with sparse $\mathbf{W}^+$ and $\mathbf{W}^-$), dominated by the gradient computation (8).

4.4. Algorithms to find the optimal embedding

We now have two ways of finding the optimal embedding. We can run the homotopy method by increasing (relatively slowly) λ from above the critical value λ_1^* (in practice, from above its upper bound u_1), minimising at each step $E(\mathbf{X};\lambda)$ over $\mathbf{X}$ and tracking the path $\mathbf{X}(\lambda)$, and stopping when the scale grows logarithmically. Other than not increasing λ too fast, no special care or ad-hoc user parameters are needed for the optimisation. This makes the algorithm almost deterministic, in that the early exploration at small scales can find and track the same, deep minimum, and in our experience produces good results, but is slower.

The second, much faster way is to select a large enough λ, fix it, and optimise there. The result then does depend on the initial $\mathbf{X}$, which can be taken random, or from the embedding of a spectral method (rescaled to match the EE scale at λ, which can be estimated in a quick, pilot run from random $\mathbf{X}$). Note that SNE implicitly sets $\lambda=1$.

A third approach, not explored here, is to minimise E subject to quadratic constraints on $\mathbf{X}$ (as in LE, LLE, etc.). Then, the solution for $\lambda=0$ is not anymore $\mathbf{X}=\mathbf{0}$, but the solution of the corresponding spectral problem, which is a better initial point. However, the now constrained optimisation is more difficult.

4.5. Out-of-sample extension

EE (like SNE and spectral methods) returns low-dimensional projections $\mathbf{X}$ only for points in the training set $\mathbf{Y}$. One way to define mappings $\mathbf{F}$ and $\mathbf{f}$ that apply to new points $\mathbf{y}$ or $\mathbf{x}$, respectively, is to fit them to $(\mathbf{Y},\mathbf{X})$ or $(\mathbf{X},\mathbf{Y})$, respectively. Although this can be made to work in practice, the result does depend on the choice of mappings, which is left to the user. Here we follow instead the more natural approach proposed in the LELVM model of Carreira-Perpiñán & Lu (2007) for LE, which returns a nonparametric mapping. The idea is, given a new point $\mathbf{y}$, to solve the original problem (i.e., to minimise the EE error E) over the unknown projection $\mathbf{x}$, keeping all other projections $\mathbf{X}$ fixed (so as not to disturb the embedding we have already obtained). The same idea applies to map a new $\mathbf{x}$ to the $\mathbf{y}$-space. Then, the error function augmented with $\mathbf{y}$ and $\mathbf{x}$ consists of the old error function $E(\mathbf{X})$ (applied to $\mathbf{X}$ and $\mathbf{Y}$) plus the new term

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

with kernels

$$w^+(\mathbf{y},\mathbf{y}_n) = \exp\big(-\tfrac{1}{2}\,\|(\mathbf{y}-\mathbf{y}_n)/\sigma\|^2\big)$$
$$w^-(\mathbf{y},\mathbf{y}_n) = \widetilde{w}_n^-\,\|\mathbf{y}-\mathbf{y}_n\|^2$$

induced from the affinity kernels that were used in the EE training (using the same neighbourhood structure). Then, we define the dimensionality reduction and reconstruction mappings as follows:

$$\mathbf{F}(\mathbf{y}) = \arg\min_{\mathbf{x}} E'(\mathbf{x},\mathbf{y}) \quad \mathbf{f}(\mathbf{x}) = \arg\min_{\mathbf{y}} E'(\mathbf{x},\mathbf{y}) \tag{13}$$

initialising $\mathbf{x}$ to the $\mathbf{x}_n$ whose $\mathbf{y}_n$ is closest to $\mathbf{y}$ in Euclidean distance, and analogously for $\mathbf{f}$. Unlike in the LELVM model, where these problems had a closed-form solution, in our case they are nonlinear optimisation problems that can be solved e.g. with gradient descent. By equating the gradient to zero we see that the minima of eq. (13) have the form of a linear combination of $\mathbf{X}$ or $\mathbf{Y}$, e.g.

$$\mathbf{x} = \sum_{n=1}^{N}\frac{w(\mathbf{y},\mathbf{y}_n)}{\sum_{n'=1}^{N} w(\mathbf{y},\mathbf{y}_{n'})}\mathbf{x}_n = \mathbf{F}(\mathbf{y})$$
$$w(\mathbf{y},\mathbf{y}_n) = w^+(\mathbf{y},\mathbf{y}_n) - \lambda w^-(\mathbf{y},\mathbf{y}_n)\exp\big(-\|\mathbf{x}-\mathbf{x}_n\|^2\big)$$

but, unlike in the LELVM model, this linear combination is not necessarily convex because the weights can be negative. Thus, the out-of-sample mappings can result in values beyond the convex hull of the data, and this allows to extrapolate to some extent.

5. Experimental Results

Fig. 2 shows the result of EE with a 2D spiral with $N=200$ points and full-graph Gaussian affinities with $\sigma=0.05$. We show all the lower and upper bounds, which imply $\lambda_1^* \in [4\cdot 10^{-5},\ 10^{-4}]$ (vertical blue lines in

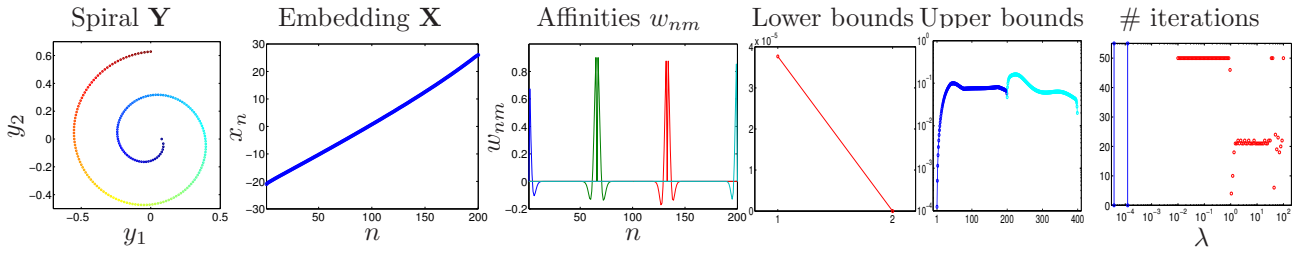

Figure 2. EE trained with homotopy with a 2D spiral.

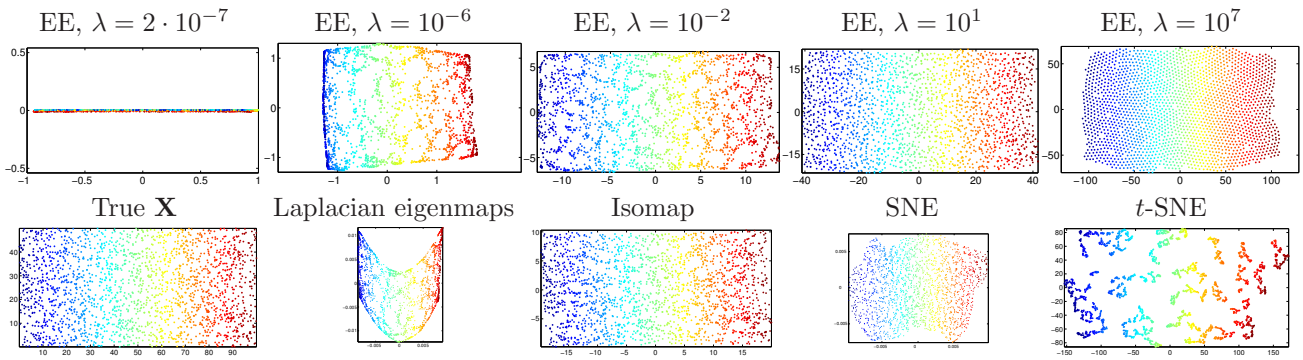

Figure 3. Swiss roll. *Top*: EE with homotopy; we show $\mathbf{X}$ for different λ. *Bottom*: true $\mathbf{X}$ and results with other methods.

the right plot). We used the homotopy method with 80 values of λ from 10^{-2} to 10^{2}. For each λ we ran the optimisation until the relative function change was less than 10^{-3} or we reached 50 iterations. The step size was 1 nearly always, 0.8 occasionally. The right plot shows that more iterations are required shortly after the λ_1^* bifurcation; occasional spikes in that plot indicate subsequent bifurcations as new minima arise and the map changes. The initial $\mathbf{X}$ do not unfold the spiral correctly, but eventually they do, and this deep minimum is tracked henceforth. As λ increases, initial local clustering and boundary effects typically associated with an LE embedding are removed and the result is a perfectly spaced sequence matching the data spacing. The initial affinities w_{nm} of eq. (9) are Gaussian, but as λ increases they develop negative lobes and adopt a Mexican-hat form (the plot shows w_{nm} for two interior and two extreme points). As λ further increases (enlarging the map and forcing points to be equidistant) w_{nm} become much more negative.

Fig. 3 shows the result of EE with a 3D Swiss roll with $N = 2\,000$ points, w_{nm}^+ as k-nearest-neighbour Gaussian affinities and $\overline{w}_{nm}^- = 1\ \forall n, m$. We set $k = 12$, $\sigma = 15$ for all methods. The bounds indicate $\lambda_1^* \in [5 \cdot 10^{-9},\ 10^{-8}]$, so we varied λ from 10^{-7} to 10^{7}. After the critical λ_1^*, $\mathbf{X}$ expands along the 1D LE solution and later on the 2D map unfolds. This small-λ solution globally unfolds the Swiss roll but shows defects similar to those of spectral methods (local clusters and gaps, boundary effects; see the LE plot). But these disappear as λ increases; $\mathbf{X}$ for $\lambda \in [10^{-1}, 10^{1}]$ is extremely similar to the true $\mathbf{X}$ (see also the result of Isomap, ideally suited to this problem). For very large λ, in the region of log-growth of the scale (see fig. 1 right), the point-separating prior dominates and the 2D arrangement tends to a round hexagonal grid (that still preserves the global structure, though). SNE attains a good map, better than LE's but worse than EE's. However, t-SNE does poorly, grouping points in local clusters that push away from each other. As noted in the introduction, t-SNE was designed to correct the map when the its dimension does not match the intrinsic one (not the case here). Initialising $\mathbf{X}$ from the true $\mathbf{X}$ produces similar results for SNE and t-SNE, indicating this is not just a bad local optimum. For SNE, perhaps better results would be obtained if

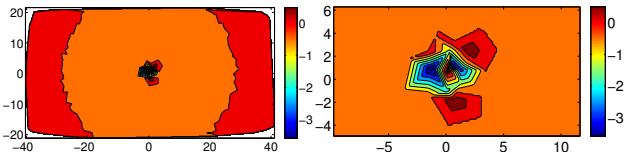

Figure 4. Affinities $w_{nm} = w_{nm}^+ - \lambda w_{nm}^- \exp\left(-\|\mathbf{x}_n - \mathbf{x}_m\|^2\right)$ learned for a point $\mathbf{x}_n$ near the centre of the Swiss roll for $\lambda = 10^{1}$ (right plot: zoom view).

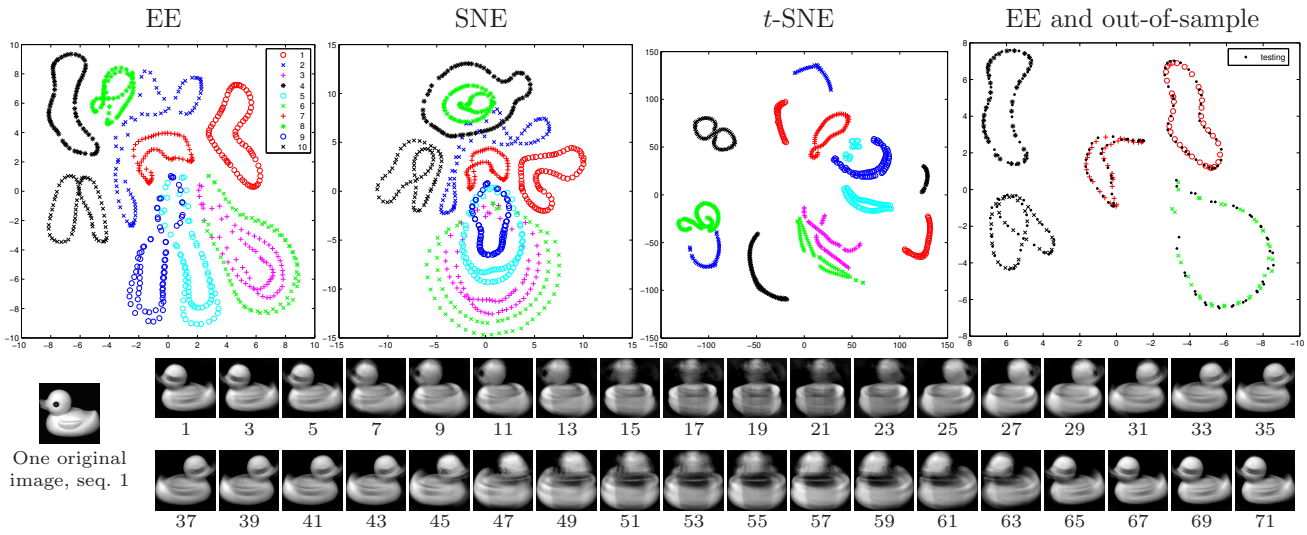

Figure 5. Results of EE, SNE and t-SNE with the COIL-20 dataset, all randomly initialised. *Right plot*: result of EE when trained on half the data, and points $\mathbf{x} = \mathbf{F}(\mathbf{y})$ ($\bullet$ marks) predicted for the other half of the data (only 5 of the sequences are shown to avoid clutter). *Below*: images $\mathbf{y} = \mathbf{f}(\mathbf{x})$ predicted for out-of-sample points in $\mathbf{x}$-space along sequence 1.

introducing a λ weight as in EE. Fig. 4 shows that the affinities w_{nm} again evolve from initially Gaussian to a 2D Mexican-hat shape, with a central positive region, negative intermediate regions and zero farther away.

Fig. 5 shows the result with the COIL-20 dataset, containing rotation sequences of 10 objects every 5 degrees, each a greyscale image of 128×128 pixels (total $N = 720$ points in $D = 16\,384$ dimensions). Thus, this contains ten 1D manifolds. We did not apply PCA to the images, and used SNE affinities with perplexity $k = 12$. We ran EE for a fixed $\lambda = 1$ from a random initial $\mathbf{X}$. The results shown for EE, SNE and t-SNE are quite robust (e.g. initialising one method from the result of another produces very similar maps). They again indicate that EE (and SNE) do a good job at both separating different clusters and capturing each sequence's 1D order. A self-intersection or a sequence that folds over itself (e.g. sequence 5) is mostly caused by quasi-symmetric COIL-20 objects that look very similar from the back or the front. t-SNE is very good at separating clusters but unfortunately it also separates parts of a cluster; most sequences appear in several pieces and folded over themselves. A further advantage of EE is that it trains faster than SNE or t-SNE.

Fig. 5 (rightmost) shows the result of training EE with half of the data (the even-numbered images in each sequence). We computed the 2D projection $\mathbf{x}_m = \mathbf{F}(\mathbf{y}_m)$ of each of the test $\mathbf{y}_m$ (odd-numbered images) with EE's out-of-sample extension. They project to their expected locations: between each even image, and in pairs for sequence 6 which folded over itself. We then mapped these $\mathbf{x}_m$ back to image space as $\mathbf{y}'_m = \mathbf{f}(\mathbf{x}_m) = \mathbf{f}(\mathbf{F}(\mathbf{y}_m))$ and achieved the reconstructions shown. Although blurred (remember we are using only 2 latent dimensions) they perfectly capture the viewpoint and general structure of the object.

6. Discussion

The intuition of Hinton & Roweis (2003) in proposing SNE was to emphasise both local and global distances through the matching of data and latent probabilities P and Q, so that e.g. making q_{nm} large when p_{nm} was small would waste some of the probability mass in Q. In our view, it is not the (normalised) probabilities that make SNE do well. We can use other (non-normalised) affinities for p_{nm} and get good results, and we have shown that the normalisation term in q_{nm}, which gives rise to the cumbersome log-sum term in E_{SNE}, is unnecessary. Rather, what makes SNE, t-SNE and EE do well is the use of Gaussian or other decaying functions to modulate the relative contributions of local vs global distances, and this is more easily achieved by EE.

EE reveals an important relation between nonlinear and spectral methods in that, at a fixed λ, EE both learns the embedding $\mathbf{X}$ and the pairwise affinities $\mathbf{W}$. We already know that putting some effort in learning good affinities can give better results than a simple functional form (e.g. Gaussian), as in the MVU method (Weinberger & Saul, 2006). However,

most work on spectral methods for dimensionality reduction and clustering and in kernel methods has focused on nonnegative, positive semidefinite affinity matrices (although nonpositive affinities do arise naturally in some constrained clustering algorithms, e.g. Lu & Carreira-Perpiñán, 2008). Our results suggest that some eigenvectors (not necessarily the extremal ones) resulting from affinity matrices that are not positive definite and that have negative entries may contain far more useful information. Remarkably, the affinities learned by EE look like Mexican-hat functions (that adapt to each point) in the λ region were the best maps arise. It is intriguing that similar Mexican-hat functions are a fundamental component in the majority of the models proposed to explain pattern formation in cortical maps, in particular the elastic net (Carreira-Perpiñán & Goodhill, 2004).

7. Conclusion

Our paper has proposed an algorithm that we think improves over SNE, producing results of similar or better quality more quickly and robustly. We have given a theoretical study of its homotopy parameter and of efficient, parameter-free, globally convergent optimisation algorithms, and an out-of-sample extension. All these ideas are directly applicable to SNE, t-SNE and earlier algorithms. Beyond this, our work has explored a new direction that we hope will spur further research: the relation of nonlinear methods such as SNE or EE with spectral methods, and the learning of affinities.

Acknowledgements

I thank Jianwu Zeng for help with the figures. Work supported by NSF CAREER award IIS–0754089.

References

Belkin, Mikhail and Niyogi, Partha. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, June 2003.

Borg, Ingwer and Groenen, Patrick. *Modern Multidimensional Scaling: Theory and Application*. Springer-Verlag, second edition, 2005.

Carreira-Perpiñán, Miguel Á. and Goodhill, Geoffrey J. Influence of lateral connections on the structure of cortical maps. *J. Neurophysiol.*, 92(5):2947–2959, November 2004.

Carreira-Perpiñán, Miguel Á. and Lu, Zhengdong. The Laplacian Eigenmaps Latent Variable Model. In *Proc. AISTATS*, pp. 59–66, San Juan, Puerto Rico, March 21–24 2007.

Carreira-Perpiñán, Miguel Á. and Lu, Zhengdong. Dimensionality reduction by unsupervised regression. In *Proc. CVPR*, Anchorage, AK, June 23–28 2008.

Cook, James, Sutskever, Ilya, Mnih, Andriy, and Hinton, Geoffrey. Visualizing similarity data with a mixture of maps. In *Proc. AISTATS*, San Juan, Puerto Rico, March 21–24 2007.

Durbin, Richard, Szeliski, Richard, and Yuille, Alan. An analysis of the elastic net approach to the traveling salesman problem. *Neural Computation*, 1(3): 348–358, Fall 1989.

Hastie, Trevor J. and Stuetzle, Werner. Principal curves. *J. Amer. Stat. Assoc.*, 84(406):502–516, June 1989.

Hinton, Geoffrey and Roweis, Sam T. Stochastic neighbor embedding. In *Proc. NIPS 2002*, pp. 857–864. MIT Press, Cambridge, MA, 2003.

Lu, Zhengdong and Carreira-Perpiñán, Miguel Á. Constrained spectral clustering through affinity propagation. In *Proc. CVPR*, Anchorage, AK, June 23–28 2008.

Nocedal, Jorge and Wright, Stephen J. *Numerical Optimization*. Springer-Verlag, second edition, 2006.

Roweis, Sam T. and Saul, Lawrence K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, December 22 2000.

Sammon, Jr., John W. A nonlinear mapping for data structure analysis. *IEEE Trans. Computers*, C–18 (5):401–409, May 1969.

Tenenbaum, Joshua B., de Silva, Vin, and Langford, John C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 22 2000.

van der Maaten, Laurens J. P. and Hinton, Geoffrey E. t-distributed stochastic neighbor embedding. *Journal of Machine Learning Research*, 9:2579–2605, November 2008.

Venna, Jarkko and Kaski, Samuel. Nonlinear dimensionality reduction as information retrieval. In *Proc. AISTATS*, San Juan, Puerto Rico, March 21–24 2007.

Weinberger, Kilian Q. and Saul, Lawrence K. Unsupervised learning of image manifolds by semidefinite programming. *Int. J. Computer Vision*, 70(1):77–90, October 2006.