Global versus local methods in nonlinear dimensionality reduction

Vin de Silva
Department of Mathematics,
Stanford University,
Stanford. CA 94305
silva@math.stanford.edu

Joshua B. Tenenbaum
Department of Brain and Cognitive Sciences,
Massachusetts Institute of Technology,
Cambridge. MA 02139
jbt@ai.mit.edu

Abstract

Recently proposed algorithms for nonlinear dimensionality reduction fall broadly into two categories which have different advantages and disadvantages: global (Isomap [1,2]), and local (Locally Linear Embedding [3], Laplacian Eigenmaps [4]). In this paper we describe variants of the Isomap algorithm which overcome two of the apparent disadvantages of the global approach.

1 Introduction

In this paper we discuss the problem of non-linear dimensionality reduction (NLDR): the task of recovering meaningful low-dimensional structures hidden in high-dimensional data. An example might be a set of pixel images of an individual's face observed under different pose and lighting conditions; the task is to identify the underlying variables (pose angles, direction of light, distance from camera, etc.) given only the high-dimensional pixel image data. In many cases of interest, the observed data are found to lie on an embedded submanifold of the high-dimensional space. The degrees of freedom along this submanifold correspond to the underlying variables. In this form, the NLDR problem is known as "manifold learning".

Classical techniques for manifold learning, such as principal components analysis (PCA) or multidimensional scaling (MDS), are designed to operate when the submanifold is embedded linearly, or almost linearly, in the observation space. More generally there is a wider class of techniques, involving iterative optimization procedures, by which unsatisfactory linear representations obtained by PCA or MDS may be "improved" towards more successful non-linear representations of the data. These techniques include GTM [9], self organising maps [10] and others.

However, Tenenbaum [1] observed that such algorithms often fail when non-linear structure cannot simply be regarded as a perturbation from a linear approximation. The iterative approach has a tendency to get stuck at locally optimal solutions that grossly misrepresent the true geometry of the situation. This is not just a theoretical issue; a simple dataset like the Swiss roll (Figure 2) will typically defeat these methods.

Recently, several entirely new approaches have been devised to address this problem. These methods combine the advantages of PCA and MDS—computational efficiency; few free

parameters; non-iterative global optimisation of a natural cost function—with the ability to disentangle the Swiss roll and other classes of nonlinear data manifold.

These algorithms come in two flavors: local and global. Local approaches (LLE [3], Laplacian Eigenmaps [4]) essentially seek to map nearby points on the manifold to nearby points in low-dimensional space. Global approaches (such as Isomap [2]) may similarly seek to map nearby points on the manifold to nearby points in low-dimensional space, but at the same time faraway points on the manifold must be mapped to faraway points in low-dimensional space.

The principal advantages of the global approach are that it tends to give a more faithful representation of the data's global structure, and that its metric-preserving properties are better understood theoretically. The local approaches have two principal advantages: (1) computational efficiency: they involve only sparse matrix computations which may yield a polynomial speedup; (2) representational capacity: they may give useful results on a broader range of manifolds, whose local geometry is close to Euclidean, but whose global geometry may not be.

In this paper we show how the global geometric approach, as implemented in Isomap, can be extended in both of these directions. The results are computational efficiency and representational capacity equal to or in excess of existing local approaches (LLE, Laplacian Eigenmaps), but with the greater stability and theoretical tractability of the global approach. Conformal Isomap (or C-Isomap) is an extension of Isomap which is capable of learning the structure of certain curved manifolds. This extension comes at the cost of making a uniform sampling assumption about the data. Landmark Isomap (or L-Isomap) is a technique for approximating a large global computation in Isomap by a much smaller set of calculations. The bulk of the work is confined to a small subset of the data, called landmark points.

The remainder of the paper is in two sections. In Section 2, we describe a perspective on manifold learning in which C-Isomap appears as the natural generalisation of Isomap. In Section 3 we derive L-Isomap from a landmark verion of classical MDS.

2 Isomap for conformal embeddings

2.1 Manifold learning and geometric invariants

We can view the problem of manifold learning as an attempt to invert a generative model for a set of observations. Let Y be a d-dimensional domain contained in the Euclidean space $\mathbf{R}^d$, and let $f : Y \to \mathbf{R}^D$ be a smooth embedding, for some $D > d$. The object of manifold learning is to recover Y and f based on a given set $\{x_i\}$ of observed data in $\mathbf{R}^D$. The observed data arise as follows. Hidden data $\{y_i\}$ are generated randomly in Y, and are then mapped by f to become the observed data, so $\{x_i = f(y_i)\}$.

The problem as stated is a little unfair (if not ill-posed). Some restriction is needed on f if we are to relate the observed geometry of the data to the structure of the hidden variables $\{y_i\}$ and Y itself. In this paper we will discuss two possibilities. The first is that we assume f to be an isometric embedding in the Riemannian sense. This means that f preserves lengths and angles at an infinitesimal scale. The other possibility we entertain is that f is a conformal embedding; so it preserves angles (but not lengths) at an infinitesimal scale. This means that at any given point y there is a scale factor $s(y) > 0$ so that very close to y, the effect of f is to magnify distances by a factor of $s(y)$. The class of conformal embeddings includes all isometric embeddings as well as many other classes of maps, including stereographic projections like the Mercator projection.

One approach to solving a manifold learning problem is to identify which aspects of the geometry of Y are invariant under the mapping f. For example, if f is an isometric embed-

ding then by definition infinitesimal distances are preserved. But more is true. The length of a path in Y is defined by integrating the infinitesimal distance metric along the path. The same is true in $f(Y)$, so f preserves path lengths. It follows that if y, z are two points in Y, then the *shortest* path between y and z lying inside Y is the same length as the shortest path between $f(y)$ and $f(z)$ along $f(Y)$. Thus geodesic distances are preserved.

The conclusion is that Y, regarded as a metric space under geodesic distance, is isometric with $f(Y)$, regarded similarly. Isomap exploits this idea by constructing the geodesic metric for $f(Y)$, at least approximately as a matrix, using the observed data alone.

To solve the conformal embedding problem, we need to identify an observable geometric invariant of conformal maps. Since conformal maps are locally isometric up to a scale factor $s(y)$, one approach is to attempt to identify or estimate $s(y)$ at each point $f(y)$ in the observed data. Then, by rescaling, we can identify the original metric structure of the data amd proceed as in Isomap. A side effect of local scaling is that the local volume of Y is scaled by a factor of $s(y)^d$. If data are generated randomly in Y, this will manifest itself by a change in the density of data points before and after applying f. In particular, if the hidden data are sampled *uniformly* on Y, then the local density of the observed data will be enough to identify the factor $s(y)$.

C-Isomap does exactly that. Under a uniform sampling assumption, if f is a conformal embedding then C-Isomap estimates the factor $s(y)$ and hence the original geometric structure of the data. In the next section, we describe the algorithms more specifically.

2.2 Isomap and C-Isomap

This is the standard Isomap procedure [2]:

1. Determine a *neighbourhood graph* G of the observed data $\{x_i\}$ in a suitable way. For example, G might contain $x_i x_j$ iff x_j is one of the k nearest neighbours of x_i (and vice versa). Alternatively, G might contain the edge $x_i x_j$ iff $|x_i - x_j| < \epsilon$, for some ϵ.
2. Compute shortest paths in the graph for all pairs of data points. Each edge $x_i x_j$ in the graph is weighted by its Euclidean length $|x_i - x_j|$, or by some other useful metric.
3. Apply MDS to the resulting shortest-path distance matrix D to find a new embedding of the data in Euclidean space, approximating Y.

The premise is that local metric information (in this case, lengths of edges $x_i x_j$ in the neighbourhood graph) is regarded as a trustworthy guide to the local metric structure in the original (latent) space. The shortest-paths computation then gives an estimate of the global metric structure, which can be fed into MDS to produce the required embedding.

It is known that Isomap converges asymptotically to the true underlying structure, given sufficient data. More precisely, a theorem of the following form is proved in [5]:

Theorem. *Let Y be sampled from a bounded convex region in $\mathbf{R}^d$, with respect to a density function $\alpha = \alpha(y)$. Let f be a C^2-smooth isometric embedding of that region in $\mathbf{R}^k$. Given $\lambda, \mu > 0$, for a suitable choice of neighborhood size parameter ϵ or k, we have*

$$1 - \lambda \leq \frac{\text{recovered distance}}{\text{original distance}} \leq 1 + \lambda$$

with probability at least $1 - \mu$, provided that the sample size is sufficiently large. [The formula is taken to hold for all pairs of points simultaneously.]

C-Isomap is a simple variation on Isomap. Specifically, we use the k-neighbours method in Step 1, and replace Step 2 with the following:

2a. Compute shortest paths in the graph for all pairs of data points. Each edge $x_i x_j$ in the graph is weighted by $|x_i - x_j| / \sqrt{M(i)M(j)}$. Here $M(i)$ is the mean distance of x_i to its k nearest neighbours.

Using similar arguments to those in [5], it is possible to prove convergence result for C-Isomap. The exact formula for the weights is not critical in the asymptotic analysis. The point is that the rescaling factor $\sqrt{M(i)M(j)}$ is an asymptotically accurate approximation to the conformal scaling factor near x_i and x_j.

Theorem. *Let Y be sampled uniformly from a bounded convex region in $\mathbf{R}^d$. Let f be a C^2-smooth conformal embedding of that region in $\mathbf{R}^N$. Given $\lambda, \mu > 0$, for a suitable choice of neighborhood size parameter k, we have*

$$1 - \lambda \leq \frac{\text{recovered distance}}{\text{original distance}} \leq 1 + \lambda$$

with probability at least $1 - \mu$, provided that the sample size is sufficiently large.

Explicit lower bounds for the sample size are much more difficult to formulate here; certainly we expect to require a larger sample than in regular Isomap to obtain good approximations. In situations where both Isomap and C-Isomap are applicable, it may be preferable to use Isomap, since it is less susceptible to local fluctuations in the sample density.

2.3 Examples

We ran C-Isomap, Isomap, MDS and LLE on two toy "fishbowl" data sets and one more realistic simulated data set. Output plots are shown in Figure 3.

Conformal fishbowl: 2000 points were generated uniformly in a circular disk and stereographically projected (thus, conformally mapped) onto a sphere. Both MDS and Isomap fail, unsurprisingly, to recognize the original disk structure of the data. C-Isomap behaves exactly as predicted, flattening the disk convincingly. LLE is just as successful.

Asymmetric fishbowl: This time, 2000 points were generated somewhat asymmetrically on a disk (using a center-offset Gaussian distribution); The purpose being to test the stability of C-Isomap and LLE in situations when the data sampling density is not perfectly uniform. MDS and Isomap behave much as with the conformal fishbowl. C-Isomap still flattens the disk, but the edges are not quite fully flattened. LLE lays out most of the disk successfully, but one sector fails to resolve correctly.

Face images: Artificial images of a face were rendered using a software package ("Poser", by Curious Labs), varying two parameters independently. In this case the parameters were left-right pose angle and distance from the camera. 128×128 color pixel images were converted into grayscale and treated as vectors in 16384-dimensional space. Ignoring perspective distortions for the closest images, there is a natural family of conformal transformations in this data set. If z is the distance variable, then transformations of the form $z \mapsto \lambda z$ are all approximately conformal, since the effect is to shrink or magnify the apparent size of each image by a constant factor. Sampling uniformly in the pose variable and logarithmically in the distance variable therfore gives a conformally uniform probability density. We generated 2000 face images in this way, spanning the range indicated by Figure 1. All four algorithms returned a two-dimensional embedding of the data. As expected, C-Isomap returns the cleanest embedding, separating the two degrees of freedom reliably along the horizontal and vertical axes. Isomap returns an embedding which narrows predictably as the face gets further away. In contrast, LLE gives an extremely distorted embedding.

Figure 1: A set of 2000 face images were randomly generated, varying independently in two parameters: distance and left-right pose. The four extreme cases are shown.

3 Isomap with landmark points

The standard Isomap algorithm tends to have bottlenecks in two places. First, one has to calculate the $N \times N$ shortest-path distance matrix D_N. The simplest algorithm is Floyd's, with complexity $O(N^3)$. This can be improved to $O(N^2 \log N)$ by implementing a version of Dijkstra's algorithm with Fibonacci heaps. After computing D_N, the subsequent MDS calculation involves an $N \times N$ symmetric eigenvalue problem. The matrix involved is full (as opposed to sparse), so this is an $O(N^3)$ problem. This is where Isomap suffers in comparison with LLE or Laplacian Eigenmaps, which reduce to a sparse symmetric eigenvalue problem.

The purpose of L-Isomap is to kill two birds with one stone. We designate n of the data points as *landmark* points, where $n \ll N$. Instead of computing D_N, we compute the $n \times N$ matrix $D_{n,N}$ of distances of each data point to the landmark points only. Then we somehow use $D_{n,N}$ (instead of D_N) to find a Euclidean embedding of the whole data set. We refer to this last step as Landmark MDS (or L-MDS). The calculation of $D_{n,N}$ by Dijsktra is $O(nN \log N)$ and our proposed algorithm for L-MDS runs in $O(n^2 N)$.

Why is it reasonable to expect L-MDS to be feasible? We can begin by applying classical MDS to the $n \times n$ landmarks-only distance matrix D_n. This provides a faithful low-dimensional embedding of the landmark points, say in $\mathbf{R}^k$. We now wish to embed the remaining points in $\mathbf{R}^k$. For each point x if we know the distances $|x - \ell|$ to each landmark point ℓ, we get n constraints on the embedding of x. If $n > k$ and the landmarks are in general position, then we have enough constraints to determine the location uniquely (if it exists).

This last step is exactly analogous to triangulating a position from exact knowledge of the distances from a small number of global positioning satellites. We will give an explicit formula and state some of its properties. For stability one generally selects a larger number of landmark points than the bare minimum ($n = k + 1$) required for a k-dimensional embedding.

3.1 The Landmark MDS procedure

Classical MDS proceeds as follows [6,7], starting with the (landmarks-only) distance matrix D_n. It is convenient to write Δ_n for the matrix of *squared* distances. The first step is to manufacture an "inner-product" matrix $B_n = -H_n \Delta_n H_n / 2$ where H_n is the *centering matrix* defined by the formula $(H_n)_{ij} = \delta_{ij} - 1/n$. Next we find the eigenvalues and eigenvectors of B_n. Write λ_i for the positive eigenvalues (labelled so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$), and $\vec{v}_i$ for the corresponding eigenvectors (written as column vectors); non-positive eigenvalues are ignored. Then for $k \leq p$ the required optimal k-dimensional embedding vectors

are given as the columns of the matrix:

$$L = \begin{bmatrix} \sqrt{\lambda_1} \cdot \vec{v}_1^T \\ \sqrt{\lambda_2} \cdot \vec{v}_2^T \\ \vdots \\ \sqrt{\lambda_k} \cdot \vec{v}_k^T \end{bmatrix}$$

The embedded vectors are automatically mean-centered, and the principal components of the embedded points are aligned with the axes, most significant first. If B_n has no negative eigenvalues, then the p-dimensional embedding is perfect; otherwise there is no exact Euclidean embedding.

For L-MDS we must now embed the remaining points in $\mathbf{R}^k$. Let Δ_x denote the column vector of squared distances between a data point x and the landmark points. It turns out that the embedding vector for x is related linearly to Δ_x. The formula is:

$$\vec{x} = \frac{1}{2} L^\dagger (\bar{\Delta}_n - \Delta_x)$$

where $\bar{\Delta}_n$ is the mean of the columns of Δ_n, and $L^\dagger$ is the pseudoinverse transpose of L, given by an explicit formula:

$$L^\dagger = \begin{bmatrix} \vec{v}_1^T / \sqrt{\lambda_1} \\ \vec{v}_2^T / \sqrt{\lambda_2} \\ \vdots \\ \vec{v}_k^T / \sqrt{\lambda_k} \end{bmatrix}$$

A full discussion of this construction will appear in [8]. We note two results here:

1. If x is actually a landmark point, then the embedding given by L-MDS is consistent with the original MDS embedding.
2. If the distance matrix $D_{n,N}$ can be represented exactly by a Euclidean configuration in $\mathbf{R}^k$, and if the landmarks are chosen so that their affine span in that configuration is k-dimensional, then L-MDS will recover that configuration exactly (up to rotation and translation).

If the original distance matrix deviates only slightly from being Euclidean, then one can argue by perturbation theory that L-MDS will give an approximately correct answer, provided that the smallest eigenvalue utilised, λ_k, is not too small. If it *is* close to zero, then $L^\dagger$ will have large norm and may overly magnify small deviations from the ideal Euclidean case. In cases where the distance matrix is highly non-Euclidean, amusing examples show that L-MDS may be a very poor approximation to doing classical MDS on the full dataset.

3.2 Example

In Figure 2, we show some of the results of testing L-Isomap on a Swiss roll data set. 2000 points were generated uniformly in a rectangle (top left) and mapped into a Swiss roll configuration in $\mathbf{R}^3$. Ordinary Isomap recovers the rectangular structure correctly provided that the neighborhood parameter is not too large (in this case $k = 8$ works). The tests show that this peformance is not significantly degraded when L-Isomap is used. For each n, we chose n landmark points at random; even down to 4 landmarks the results are excellent.

In contrast, the output of LLE is quite unstable under changes in its sparseness parameter k (neighborhood size). In fairness, k is really a topological parameter and only incidentally a sparseness parameter. In L-Isomap, these two roles are separately fulfilled by k and n.

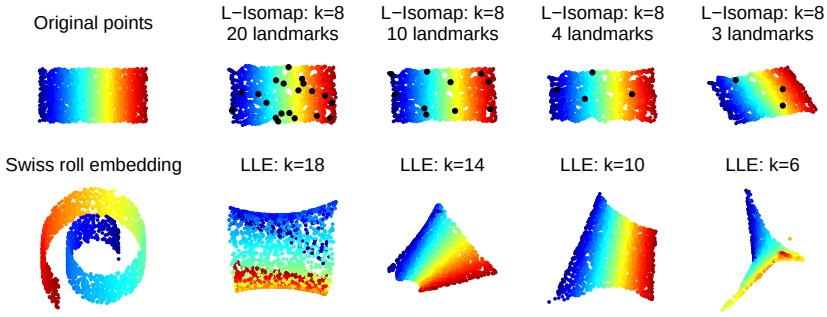

Figure 2: L-Isomap is stable over a wide range of values for the sparseness parameter n (the number of landmarks). Results from LLE are shown for comparision.

4 Conclusion

Local approaches to nonlinear dimensionality reduction such as LLE or Laplacian Eigenmaps have two principal advantages over a global approach such as Isomap: they tolerate a certain amount of curvature and they lead naturally to a sparse eigenvalue problem. However, neither curvature tolerance nor computational sparsity are explicitly part of the formulation of the local approaches; these features emerge as byproducts of the goal of trying to preserve only the data's local geometric structure. Because they are not explicit goals but only convenient byproducts, they are not in fact reliable features of the local approach. The conformal invariance of LLE can fail in sometimes surprising ways, and the computational sparsity is not tunable independently of the topological sparsity of the manifold. In contrast, we have presented two extensions to Isomap that are explicitly designed to remove a well-characterized form of curvature and to exploit the computational sparsity intrinsic to low-dimensional manifolds. We have analyzed the algorithmics of both extensions, proven the conditions under which they return accurate results, and demonstrated their success on challenging data sets.

Acknowledgments

This work was supported in part by NSF grant DMS-0101364. The authors wish to thank Lauren Schmidt for her considerable help in generating the "Tom" image dataset, using Curious Labs' "Poser" software.

References

[1] Tenenbaum, J.B. (1998) Mapping a manifold of perceptual observations. In M.I. Jordan, M.J. Kearns & S.A. Solla (eds.), *Advances in Neural Information Processing Systems 10*: Cambridge, MA: MIT Press.

[2] Tenenbaum, J.B., de Silva, V. & Langford, J.C (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* **290**: 2319–2323.

[3] Roweis, S. & Saul, L. (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**: 2323–2326.

[4] Belkin, M. & Niyogi, P. (2002) Laplacian eigenmaps and spectral techniques for embedding and clustering. In T.G. Dietterich, S. Becker and Z. Ghahramani (eds.), *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press.

[5] Bernstein, M., de Silva, V., Langford, J.C. & Tenenbaum, J.B. (December 2000) Graph approximations to geodesics on embedded manifolds. Preprint may be downloaded at the URL:

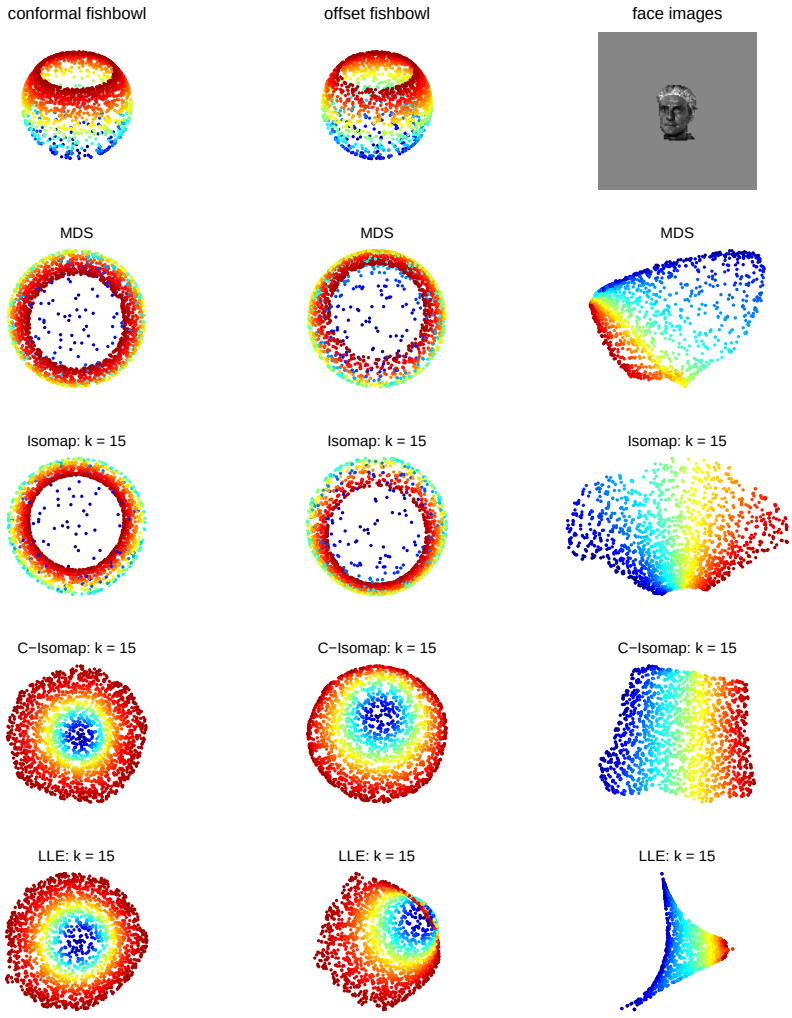

Figure 3: Four dimensionality reduction algorithms (MDS, Isomap, C-Isomap, and LLE) are applied to two toy datasets and a more complex data manifold of face images.

`http://isomap.stanford.edu/BdSLT.pdf`

[6] Torgerson, W.S. (1958) *Theory and Methods of Scaling.* New York: Wiley.

[7] Cox, T.F & Cox M.A.A (1994) *Multidimensional Scaling.* London: Chapman & Hall.

[8] de Silva, V., Tenenbaum, J.B. & Steyvers, M. (in preparation) Sparse multidimensional scaling using landmark points.

[9] Bishop, C., Svensen, M. & Williams, C. (1998) GTM: The generative topographic mapping. *Neural Computation* **10(1)**.

[10] Kohonen, T. (1984) *Self Organisation and Associative Memory* Berlin: Springer-Verlag.