# Nonlinear Manifold Learning
# 6.454 Summary

Alexander Ihler
*ihler@mit.edu*

October 6, 2003

**Abstract**

Manifold learning is the process of estimating a low-dimensional structure which underlies a collection of high-dimensional data. Here we review two popular methods for nonlinear dimensionality reduction, locally linear embedding (LLE, [1]) and IsoMap [2]. We also discuss their roots in principal component analysis and multidimensional scaling, and provide a brief comparison of the underlying assumptions, strengths, and weaknesses of each algorithm.

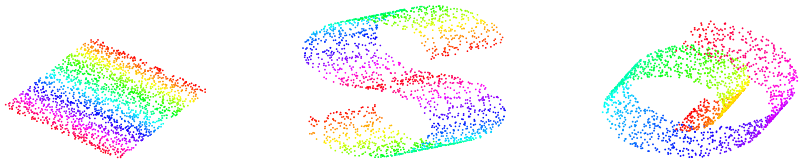
## 1 Introduction

Finding low-dimensional representations of high-dimensional data is a common problem in science and engineering. High-dimensional observations result from any of a number of data collection methods: images, spectral representations, or simply sets of associated measurements. Often however, these observations result from changes to only a few degrees of freedom, lending a predictable structure to the high-dimensional data. Algorithms for finding this underlying structure within the data are collectively referred to by the term "manifold learning".

Finding the structure behind the data may be important for a number of reasons. One possible application is data visualization. It is difficult to display and understand relationships in dimensions higher than two or three, making the use of low-dimensional transformations of the data appealing. However, such transformations must preserve (if not clarify) the underlying structure and relationships in order to be of use. It may also be desirable to measure how distant two observation pairs are in the underlying (generative) parameter space, as this is a more meaningful measure of dissimilarity than distance in the observation space.

Figure 1 shows a two-dimensional data set, embedded in three dimensions (and re-projected for printed visualization) in three different ways: a linear embedding (plane) (a), an S-shape (b), and a "Swiss roll" (c). The purpose of (nonlinear) manifold learning is to recover, in all three cases, the true dimensionality of the data and sample locations which are consistent with that geometry.

We begin our discussion of nonlinear manifold learning with a brief overview of a linear method, principal component analysis. We also present the "multidimensional



(a) (b) (c)

Figure 1: Two-dimensional manifolds embedded in three dimensions. (a) Linear embedding, (b) S-curve, (c) Swiss roll

scaling" (MDS) problem, and its solution in the classic case (principal coordinate analysis). However, these methods are insufficient for nonlinear manifolds. We then discuss two methods for estimating nonlinear embeddings. The first, locally linear embedding (LLE), finds a neighbor-based local representation of each point and then finds a low-dimensional representation with a similar configuration. The second, IsoMap, is a more direct extension to MDS, relying on the classic MDS solution but substituting an alternate estimate of distance.

## 2 Principal Components Analysis

Principal component analysis is one of the classic linear methods of dimensionality reduction [3]. Given data $X = [x_1, \ldots, x_n]$, where each $x_i$ has dimension $p$ (so that $X$ is $p \times n$), we would like to find a linear subspace of dimension $q$ such that the projected values of the $x_i$ (denoted $\hat{x}_i$) have minimal squared error. Although in this formulation the projected data $\hat{x}_i$ are of the same dimension $p$, they may then be represented as values in $\mathbb{R}^q$ (a coefficients for a set of basis vectors). We denote this lower-dimensional representation $Y$ (as $q \times n$ matrix).

It is easy to show through an orthogonality argument that minimal quadratic error in the residual is equivalent to maximizing the variance of the projected data $\hat{x}_i$; this goal can be achieved using a procedure such as the following.

We begin by subtracting the mean $\frac{1}{N}\sum_i x_i$ from each sample $x_i$; this may be accomplished by right-multiplying by the so-called "centering" matrix $J$

$$J = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}' \tag{1}$$

(where $I_n$ is the $n \times n$ identity matrix). The covariance matrix of $X$ is then given by

$$\Sigma = XJJ'X' \tag{2}$$

The subspace with maximal variance can be found easily by the eigenstructure of $\Sigma$; its (ordered) basis vectors are called the principal components. Let $\Lambda = diag(\lambda_1, \ldots \lambda_p)$ be the eigenvalues of $\Sigma$ ordered from largest to smallest, with $V_p = [v_1, \ldots, v_p]$ their corresponding eigenvectors. The location of the points $X$ in the newly defined $q$-dimensional space (where $q \leq p$) is then given by $Y = V_q' XJ$, where $V_q = [v_1, \ldots, v_q]$ are the top $q$ eigenvectors found above.

However, it may be undesirable to restrict ourselves to a linear transformation. After all, we are simply looking for a lower-dimensional representation which preserves relationships among the data, and frequently (as in Figure 1) these relationships are nonlinear. But, this begs the question of what relationships should be preserved. One simple criterion is to place some cost on distortions of the inter-sample distances; this type of criterion has received considerable attention under the name *multidimensional scaling* (MDS).

# 3 Multidimensional Scaling

Multidimensional scaling denotes the problem of finding a $q$-dimensional set of points $Y$ which are most closely consistent (under some cost function) with a measured set of "dissimilarities" $D$ (also sometimes called a "pre-distance matrix"). This problem garnered attention in the field of psychometry, by investigators who wished to build a model of the perception of various stimuli (for example, colors) using the responses of subjects to dictate dissimilarity between stimuli. Note that in psychometry, these measurements may not form a metric space (perhaps only construing an ordering of the comparisons). Incorporating this type of relationship entails more complex analysis; we ignore this subtlety here and concentrate solely on the case of Euclidean measurement observations. A nice tutorial on MDS, including a historical view of its origins in psychometry is given by [4].

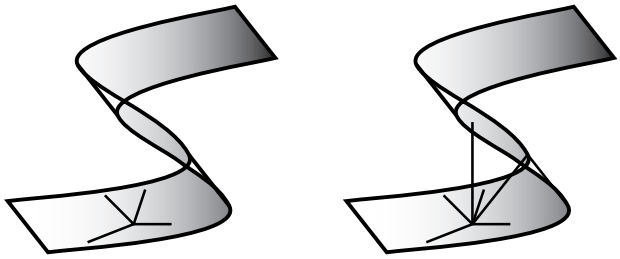
As stated, MDS is the problem of finding a low-dimensional representation which preserves the distances $D$ according to some cost $\rho(D, \hat{D})$, where $\hat{D}$ denotes the distances between the points $Y$ in the low-dimensional space. Note that in this formulation MDS places no requirements on the form or complexity of the transformation from $\mathbb{R}^p$ to $\mathbb{R}^q$, *only* on preserving pairwise relationships between the data.

For a general cost function $\rho$, this is a difficult nonlinear optimization problem; various selections of $\rho$ have given rise to a wealth of research. One very convenient form for $\rho$ (called the "STRAIN" criterion) is given by

$$\rho_{STRAIN}(D, \hat{D}) = \|J'(D^2 - \hat{D}^2)J\|_F^2 \tag{3}$$

where $D^2$ is the matrix of squared-distances (element-wise product), $\|\cdot\|_F^2$ is the (squared) Frobenius norm, and $J$ is the centering matrix defined in Section 2. The intuition behind this criterion (besides having a convenient solution) is that we would like to match *variations* in distance, rather than the values themselves; this makes the criterion invariant to, say, adding a constant to all distances, or to all distances from a given datum, potentially useful given the perceptual nature of the dissimilarity values from psychometry.

The STRAIN metric is so pervasive in MDS that it is also known as *classical MDS*; this is perhaps because it admits a convenient, closed-form solution given by the eigenstructure of $-\frac{1}{2}J'D^2J$ in the same way as Section 2. Namely, the top $q$ eigenvectors $[w_1, \ldots, w_q]$ capture the largest components of variation in $J'D^2J$, and thus give the ($q$-dimensional) coordinates of the solution $Y = [w_1, \ldots, w_q]'$. It is perhaps not surprising given the similarity of the solution that this technique should be closely related to that of principal component analysis, discussed previously.



(a) (b)

Figure 2: Euclidean distance versus geodesic: While local distances (a) are approximately equal to their Euclidean counterparts, Euclidean distances between more distant points (b) do not respect the geometry of the manifold.

Indeed, this is the case; the STRAIN metric can be shown to be a dual formulation of PCA [5] in that both recover points with the same inter-sample distance structure. This is easily seen by the fact that, again letting $X$ denote the (perhaps unknown) sample locations in $p$-dimensional space, we have $-.5J'D^2J = J'X'XJ$ (since $D_{ij}^2 = \|x_i - x_j\|^2 = \|x_i\|^2 + \|x_j\|^2 - 2x_i'x_j$ and the "double centering" operation subtracts the row and column averages, which if the $x_i$ are zero-mean are $\|x_i\|^2$ and $\|x_j\|^2$ respectively). This means that given any eigenvector $v_i$ of $\Sigma = XJJ'X'$ (as defined in Section 2), we have a corresponding eigenvector $w_i = J'X'v_i$ for $J'X'XJ$. Thus both methods resolve the same $q$-dimensional point locations $Y = W_q' = V_q'XJ$.

Other criteria for MDS have also been studied, for example the "STRESS" and "SSTRESS" criteria given by

$$\rho_{STRESS}(D, \hat{D}) = \|D - \hat{D}\|_F^2 \tag{4}$$

and

$$\rho_{SSTRESS}(D, \hat{D}) = \|D^2 - \hat{D}^2\|_F^2 \tag{5}$$

However, available algorithms for minimizing these cost functions lack the same globally optimal convergence properties that accompany STRAIN.

The main problem in applying any of the above formulations of MDS to nonlinear manifold learning however, is their direct use of and reconstruction based on Euclidean distance measurements. The reason for this is simple — the three criteria above treat all pairwise Euclidean distance measurements equally; yet for a manifold which has been nonlinearly embedded in a higher-dimensional space, many of these distances do not respect the topology of the manifold and are thus unsuitable for use in determining the nonlinear relationship.

We investigate two methods which attempt to address this issue in differing ways. Both rely on the fact that the surface of any manifold may be *locally* approximated by a linear (tangent) subspace, in much the same way that a function may be locally approximated using its derivative (i.e. as a first-order Taylor expansion). Thus a set of local distance measurements (Figure 2(a)) are regarded as trustworthy, while longer-range relationships (Figure 2(b)) are discarded. The first method, *locally linear embedding*,

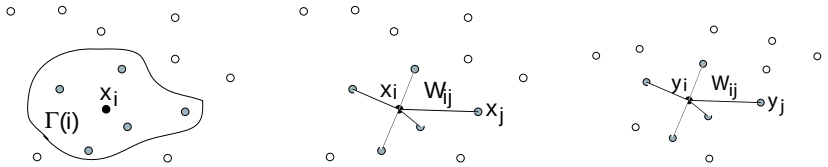



Figure 3: Locally Linear Embedding: (a) Choose a neighborhood for each point $i$, (b) find weights to reconstruct $x_i$ given its neighbors, and (c) find new lower-dimensional points $y_i$ observing these relationships.

creates and solves a set of coupled quadratic optimizations based on the locations of each point's neighbors; the second (*IsoMap*) uses local distances to approximate the true curve length along the manifold (called the geodesic) and obtain new estimates of all pairwise distances, then solves using classical MDS. We present each technique, then conclude with some remarks comparing and contrasting the two.

# 4 Locally Linear Embedding

Locally linear embedding [1], or LLE, proposes to use the local linearity of the manifold to find a weight-based representation of each point using its neighbors, characterizing the local relative positioning of each neighborhood in $\mathbb{R}^p$. Then, using this local parameterization one can look for a new set of points in a lower dimension $q$ which preserves, as closely as possible, the same relative positioning information.

First, solve for the weights which best characterize the points' relationship in $\mathbb{R}^p$:

$$\tilde{W} = \arg\min_{W} \sum_{i=1}^{n} \left\| x_i - \sum_{j \in \Gamma(i)} W_{ij} x_j \right\|^2 \qquad \text{such that} \qquad \forall i \sum_j W_{ij} = 1 \tag{6}$$

where $\Gamma(i)$ is the neighboring points of $x_i$, defined either by the $k$-nearest-neighbors or by some local sphere of radius $\epsilon$ around $x_i$. The size of this neighborhood is a compromise — it must be large enough to allow for good reconstruction of the points (contain at least $q+1$ points), but small enough for the data manifold to have little or no curvature. Defining $w_i$ to be the $i^{th}$ row of $W$ and the matrix of local difference vectors as $\Delta_i = [x_i - x_{j_1}, \ldots, x_i - x_{j_{\Gamma(i)}}]$ (where $j$ indexes the neighbors of $x_i$), Equation (6) may be rewritten in terms of each sample $x_i$:

$$\arg\min_{w_i} w_i' \Delta_i' \Delta_i w_i \qquad \text{such that} \qquad \sum_j w_{ij} = 1 \tag{7}$$

In practice $w_i$ may be found by solving

$$\Delta_i' \Delta_i w_i = \mathbf{1} \tag{8}$$

and normalizing $w_i$ to sum to unity [6].

This weight-based representation has several desirable invariances: first, it is invariant to any local rotation or scaling of $x_i$ and its neighbors (due to the linear relationship of (6)). Additionally, the normalization requirement on $w_i$ adds invariance to translation of $x_i$ and its neighbors (since $\sum_j W_{ij}(x_j + \alpha) = x_i + \sum_j W_{ij}\alpha = x_i + \alpha$). This means that LLE is capable of modeling an arbitrary nonlinear embedding function so long as it is smooth; more precisely, the mapping preserves angle and scale within each *local* neighborhood.

Having solved for the optimal weights which capture the local structure (in a mean-squared sense) at each point, we attempt to find new locations which approximate those relationships. This too can be done in closed form, by minimizing the same quadratic cost function as (6) for the new data locations:

$$\tilde{Y} = \arg\min_{Y} \sum_{i=1}^{n} \left\| y_i - \sum_{j \in \Gamma(i)} \tilde{W}_{ij} y_j \right\|^2 \quad \text{s.t.} \quad Y\mathbf{1} = \mathbf{0}, \quad YY' = I_n \tag{9}$$

(where the conditions have been added to make the problem well-posed). This can alternately be written as the quadratic form

$$\tilde{Y} = \arg\min_{Y} Y'(I - \tilde{W})'(I - \tilde{W})Y \quad \text{s.t.} \quad Y\mathbf{1} = \mathbf{0}, \quad YY' = I_n \tag{10}$$

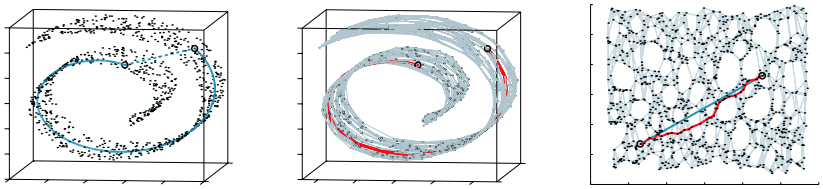
where $\tilde{W}$ is the (sparse) $n \times n$ matrix of optimal weights. This quadratic form can be composed as an eigenvector problem (similar to that of Section 2 but finding the minimum variance subspace rather than the maximum). It has the trivial eigenvector $\mathbf{1}$ (with eigenvalue zero) induced by the translational invariance resulting from the weight-selection procedure; discarding this, the remaining $q$ eigenvectors with smallest eigenvalues define the best $q$-dimensional fit.

It should also be noted that a solution can be found for all values of $q$ simultaneously; the best 1-dimensional embedding is simply the first coordinate of the best 2-dimensional embedding, and so forth.

# 5 IsoMap

As stated previously, one of the major problems with classical MDS was its use of distances calculated in the high-dimensional Euclidean space. As illustrated in Figure 2 the local Euclidean distances (a) are approximately correct, but the Euclidean distances between curved regions (b) do not respect this geometry. Thus, inclusion of those distances in our optimization means that classical MDS fails to recover the planar nature of the data.

IsoMap [2] works on a simple principle – that, given a set of distances $D$, classical MDS recovers a set of locations which best approximate (by the measure given in Section 3) those distances in a Euclidean $\mathbb{R}^q$ space. However, for a nonlinearly embedded manifold, the distances $D$ we should use are *not* the Euclidean distances between each data pair, but rather the shortest curve length which respects the geometry of the data (the geodesic). Thus, to recover a geometry using MDS we need two things: first, the value of the true (geodesic) distance measurements, and secondly, for the geodesic distance in $\mathbb{R}^q$ to be the same as Euclidean distance.



(a) (b) (c)

Figure 4: Estimating geodesics with shortest-paths: (a) in order to find the true geodesic length (solid) rather than the naive Euclidean distance (dashed), we approximate by a shortest-path length traversing trusted, local distance measurements. This is shown in the original space (b) and "unrolled" (c). Images from [2].

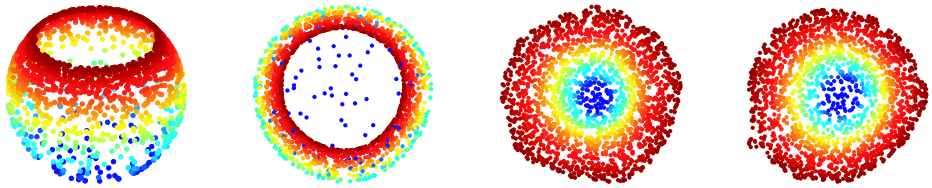
The second condition is satisfied if our low-dimensional data is located on a bounded convex subset of $\mathbb{R}^q$. To satisfy the first condition, we will estimate the geodesic curve length between distant points by assuming that our function is an isomorphism (i.e. the transformation into $\mathbb{R}^p$ preserves both angle and distance). This estimation is done by (again) appealing to the local linearity of the manifold — if our data are sufficiently dense, then there is some local neighborhood (again, either defined by the $k$-nearest neighbors or by some $\epsilon$-ball) in which the geodesic distance is well-approximated by the naive Euclidean distance in $\mathbb{R}^p$. Taking these local distances as trusted, farther distances may be approximated by finding the length of the shortest path along trusted edges; it can be shown [7] that as the number of data $n$ increases this estimate converges to the true geodesic distances. An illustration of this estimation for the Swiss roll example is shown in Figure 4.

The computational burden of this operation (naively $\mathcal{O}(n^3)$, somewhat faster with more sophisticated data structures) may be reduced by using only a subset of the points (called *landmark points*) for the classical MDS embedding step [7]. Choosing only $m$ such landmark points means that fewer pairwise geodesic distances must be computed (naively requiring $\mathcal{O}(m*n^2)$ computation), and the size of the matrix used for principal coordinate analysis is similarly reduced, requiring only $\mathcal{O}(m^2 n)$ computation. Again, the best embedding for all values of $q$ may be obtained simultaneously by computing all eigenvectors.

Additionally, IsoMap may be extended to find *conformal* maps, rather than simply isomorphic ones [7]; this adds an invariance to local scale changes much like that of LLE. (A conformal map preserves angle but not distance; it is a strictly wider class of functions than isomorphisms.) This extension is performed by further assuming that the samples are drawn *uniformly* from the convex subset of $\mathbb{R}^q$; the added assumption of uniformity enables estimation of the (local) distance distortion induced by a conformal mapping using the density of the points in a neighborhood of $R^p$.

# 6 Comparison of LLE and IsoMap

LLE and IsoMap are quite similar in goals and assumptions, and differ in execution. Both make use of the fact that the data manifold is locally linear in nature, and assume



(a) (b) (c) (d)

Figure 5: Conformal mappings: data generated on a plane and conformally warped to a fish-bowl shape (a); note the dense sampling around the rim. IsoMap (b) fails to recover the geometry due to its violated assumptions; Conformal IsoMap (c) and LLE (d) both recover the original data. Images from [7].

sufficient data that a local neighborhood of samples around each point can be found which satisfies this assumption. Both methods are sensitive to these assumptions; if the neighborhood size is chosen to be too large or the space is too sparsely sampled both methods break down (often producing a "folded" appearance).

IsoMap makes several further assumptions – that the data is actually located on a convex region of $\mathbb{R}^q$ (for some $q$) and that it was embedded in $\mathbb{R}^p$ by some isometric transformation. These assumptions allow it to estimate the true pairwise distances in the original $\mathbb{R}^q$-space and thus utilize the closed-form solution of classical MDS to obtain the embedding.

LLE, on the other hand, makes no attempt to estimate the true geodesics; it simply assumes that a weighted best-fit of each point's neighbors is sufficient to describe the local relationship of the points. Again this leads to an efficient closed-form solution to the optimization problem (over a wider space of embedding functions than traditional IsoMap), but with differing failure modes than IsoMap.

In particular, IsoMap's problems arise when its additional assumptions are violated. For example, when the mapping is conformal and not isomorphic (and this fact is not accounted for); a fishbowl example (Figure 5) illustrates Isomap's inability to recover the disc-shape which generated the data. On this example LLE succeeds, due to the invariance of the weight-based representation to local changes in scale. Alternately, when the low-dimensional points do not form a convex shape in $\mathbb{R}^q$, IsoMap will introduce distortions as it attempts to match the Euclidean (rather than geodesic) distance in $\mathbb{R}^q$ to the estimated geodesic distance in $\mathbb{R}^p$.

LLE's problems are of a different bent. Because it only makes use of the concept that two points are "near", it has no additional penalty for placing points which are not nearby in the original ($\mathbb{R}^p$) space as neighbors in the embedded ($\mathbb{R}^q$) space. This makes it resistant to the type of errors occurring in IsoMap on non-convex sets, but more susceptible to placing faraway points nearby to each other. This leads to observed "spider-web" behaviors (long, thin point sets) or a "folding" of points in the embedded coordinates (a subset of points have locations which have been reflected across some axis). Perhaps another consequence of this is an increased difficulty in making concrete when the resulting embedded coordinates recover the true geometry of the data. In fact, fixing this "too-local" cost function has been one of the thrusts of more recent research

(next week's topic).

# 7 Conclusions

Both IsoMap and LLE provide interesting alternatives to linear embedding approaches, showing considerable improvement when the underlying geometry of the the high-dimensional data is complex. Both make use of a local linearity assumption for the data, and require a sufficiently large number of samples for this assumption to be satisfied. However, the differences in underlying methodologies and additional assumptions lead to differences in performance between the two algorithms under non-ideal conditions. As we shall see next week, more recent algorithms which take more probabilistic interpretations may both improve performance and increase our understanding of the utility and difficulties in applying embedding methods.

# References


[1] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, December 2000.

[2] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), December 2000.

[3] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[4] M. W. Trosset. The formulation and solution of multidimensional scaling problems. Technical Report TR93-55, Rice University, 1993.

[5] J. C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3/4):325–338, December 1966.

[6] L. Saul and S. Roweis. Think globally, fit locally: Unsupervised learning in low-dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, June 2003.

[7] V. De Silva and J. B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In *Neural Information Processing Systems 16*. MIT Press, 2002.