A Global Geometric Framework for Nonlinear Dimensionality Reduction

Joshua B. Tenenbaum, Vin de Silva, John C. Langford

Scientists working with large volumes of high-dimensional data, such as global climate patterns, stellar spectra, or human gene distributions, regularly confront the problem of dimensionality reduction: finding meaningful low-dimensional structures hidden in their high-dimensional observations. The human brain confronts the same problem in everyday perception, extracting from its high-dimensional sensory inputs—30,000 auditory nerve fibers or 106 optic nerve fibers—a manageably small number of perceptually relevant features. Here we describe an approach to solving dimensionality reduction problems that uses easily measured local metric information to learn the underlying global geometry of a data set. Unlike classical techniques such as principal component analysis (PCA) and multidimensional scaling (MDS), our approach is capable of discovering the nonlinear degrees of freedom that underlie complex natural observations, such as human handwriting or images of a face under different viewing conditions. In contrast to previous algorithms for nonlinear dimensionality reduction, ours efficiently computes a globally optimal solution, and, for an important class of data manifolds, is guaranteed to converge asymptotically to the true structure.

A canonical problem in dimensionality reduction from the domain of visual perception is illustrated in Fig. 1A. The input consists of many images of a person's face observed under different pose and lighting conditions, in no particular order. These images can be under different pose and lighting conditions, many images of a person's face observed is illustrated in Fig. 1A. The input consists of production from the domain of visual perception. A canonical problem in dimensionality reduction, ours efficiently computes a globally optimal solution, and, for an important class of data manifolds, is guaranteed to converge asymptotically to the true structure.

The classical techniques for dimensionality reduction, PCA and MDS, are simple to implement, efficiently computable, and guaranteed to discover the true structure of data lying on or near a linear subspace of the high-dimensional input space. PCA finds a low-dimensional embedding of the data points that best preserves their variance as measured in the high-dimensional input space. Classical MDS finds an embedding that preserves the interpoint distances, equivalent to PCA when those distances are Euclidean. However, many data sets contain essential nonlinear structures that are invisible to PCA and MDS—computational efficiency, global optimality, and asymptotic convergence guarantees—with the flexibility to learn a broad class of nonlinear manifolds. Figure 3A illustrates the challenge of nonlinearity with data lying on a two-dimensional “Swiss roll”: points far apart on the underlying manifold, as measured by their geodesic, or shortest path, distances, may appear deceptively close in the high-dimensional input space, as measured by their straight-line Euclidean distance. Only the geodesic distances reflect the true low-dimensional geometry of the manifold, but PCA and MDS effectively see just the Euclidean structure; thus, they fail to detect the intrinsic two-dimensionality (Fig. 2B).

Our approach builds on classical MDS but seeks to preserve the intrinsic geometry of the data, as captured in the geodesic manifold distances between all pairs of data points. The crux is estimating the geodesic distance between faraway points, given only input-space distances. For neighboring points, input-space distance provides a good approxima-
tion to geodesic distance. For faraway points, geodesic distance can be approximated by adding up a sequence of “short hops” between neighboring points. These approximations are computed efficiently by finding shortest paths in a graph with edges connecting neighboring data points.

The complete isometric feature mapping, or Isomap, algorithm has three steps, which are detailed in Table 1. The first step determines which points are neighbors on the manifold $M$, based on the distances $d_M(i,j)$ between pairs of points $i,j$ in the input space $X$. Two simple methods are to connect each point to all points within some fixed radius $\varepsilon$, or to all of its $K$ nearest neighbors (15). These neighborhood relations are represented as a weighted graph $G$ over the data points, with edges of weight $d_G(i,j)$ between neighboring points (Fig. 3B).

In its second step, Isomap estimates the geodesic distances $d_G(i,j)$ between all pairs of points on the manifold $M$ by computing their shortest path distances $d_G(i,j)$ in the graph $G$. One simple algorithm (16) for finding shortest paths is given in Table 1.

The final step applies classical MDS to the matrix of graph distances $D_G = \{d_G(i,j)\}$, constructing an embedding of the data in a $d$-dimensional Euclidean space $Y$ that best preserves the manifold’s estimated intrinsic geometry (Fig. 3C). The coordinate vectors $y_j$ for points in $Y$ are chosen to minimize the cost function

$$E = \|\tau(D_G) - \tau(D_G)\|_2,$$

where $D_G$ denotes the matrix of Euclidean distances $\{d_G(i,j) = \|x_i - x_j\|_2\}$ and $\|A\|_2$ the $L^2$ matrix norm $\sum_{i,j} A^2_{ij}$. The $\tau$ operator

Fig. 1. (A) A canonical dimensionality reduction problem from visual perception. The input consists of a sequence of 4096-dimensional vectors, representing the brightness values of 64 pixel by 64 pixel images of a face rendered with different poses and lighting directions. Applied to $N = 698$ raw images, Isomap ($K = 6$) learns a three-dimensional embedding of the data’s intrinsic geometric structure. A two-dimensional projection is shown, with a sample of the original input images (red circles) superimposed on all the data points (blue) and horizontal sliders (under the images) representing the third dimension. Each coordinate axis of the embedding correlates highly with one degree of freedom underlying the original data: left-right pose ($x$ axis, $R = 0.99$), up-down pose ($y$ axis, $R = 0.90$), and lighting direction (slider position, $R = 0.92$). The input-space distances $d_M(i,j)$ given to Isomap were Euclidean distances between the 4096-dimensional image vectors. (B) Isomap applied to $N = 1000$ handwritten “2”s from the MNIST database (40). The two most significant dimensions in the Isomap embedding, shown here, articulate the major features of the “2”: bottom loop ($x$ axis) and top arch ($y$ axis). Input-space distances $d_M(i,j)$ were measured by tangent distance, a metric designed to capture the invariances relevant in handwriting recognition (41). Here we used ε-Isomap (with $\varepsilon = 4.2$) because we did not expect a constant dimensionality to hold over the whole data set; consistent with this, Isomap finds several tendrils projecting from the higher dimensional mass of data and representing successive exaggerations of an extra stroke or ornament in the digit.
Fig. 2. The residual variance of PCA (open triangles), MDS (open circles in [A] through (C); open circles in (D)], and Isomap (filled circles) on four data sets (42). (A) Face images varying in pose and illumination (Fig. 1A). (B) Swiss roll data (Fig. 3). (C) Hand images varying in finger extension and wrist rotation (20). (D) Handwritten "2"s (Fig. 1B). In all cases, residual variance decreases as the dimensionality d is increased. The intrinsic dimensionality of the data can be estimated by looking for the "elbow" at which this curve ceases to decrease significantly with added dimensions. Arrows mark the true or approximate dimensionality, when known. Note the tendency of PCA and MDS to overestimate the dimensionality, in contrast to Isomap.

Fig. 3. The "Swiss roll" data set, illustrating how Isomap exploits geodesic paths for nonlinear dimensionality reduction. (A) For two arbitrary points (circled) on a nonlinear manifold, their Euclidean distance in the high-dimensional input space (length of dashed line) may not accurately reflect their intrinsic similarity, as measured by geodesic distance along the low-dimensional manifold (length of solid curve). (B) The neighborhood graph G constructed in step one of Isomap (with K = 7 and N = 1000 data points) allows an approximation (red segments) to the true geodesic path to be computed efficiently in step two, as the shortest path in G. (C) The two-dimensional embedding recovered by Isomap in step three, which best preserves the shortest path distances in the neighborhood graph (overlaid). Straight lines in the embedding (blue) now represent simpler and cleaner approximations to the true geodesic paths than do the corresponding graph paths (red).
The Isomap algorithm takes as input the distances $d_{ij}$ between all pairs $i, j$ from $N$ data points in the high-dimensional input space $X$, measured either in the standard Euclidean metric (as in Fig. 1A) or in some domain-specific metric (as in Fig. 1B). The algorithm outputs coordinate vectors $y_i$ in a $d$-dimensional Euclidean space $Y$ that (according to Eq. 1) best represent the intrinsic geometry of the data. The only free parameter ($\epsilon$ or $K$) appears in Step 1.

### Table 1

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Construct neighborhood graph</td>
</tr>
<tr>
<td>2</td>
<td>Compute shortest paths</td>
</tr>
<tr>
<td>3</td>
<td>Construct $d$-dimensional embedding</td>
</tr>
</tbody>
</table>

#### References and Notes
15. The scale-invariant $K$ parameter is typically easier to set than $\epsilon$, but may yield misleading results when the local dimensionality varies across the data set. When available, additional constraints such as the temporal ordering of observations may also help to determine neighbors. In earlier work (36) we explored a more complex method (37), which required an order of magnitude more data and did not support the theoretical performance guarantees we provide here for $\epsilon$- and $K$-isomap.

#### Footnotes
- "2322" is a page number, "2020 VOL 290" is a volume and issue number, and "www.sciencemag.org" is the website of the journal Science.

---

**Fig. 4.** Interpolations along straight lines in the Isomap coordinate space (analogous to the blue line in Fig. 3C) implement perceptually natural but highly nonlinear "morphs" of the corresponding high-dimensional observations (43) by transforming them approximately along geodesic paths (analogous to the solid curve in Fig. 3A). (A) Interpolations in a three-dimensional embedding of face images (Fig. 1A). (B) Interpolations in a four-dimensional embedding of hand images (20) appear as natural hand movements when viewed in quick succession, even though no such motions occurred in the observed data. (C) Interpolations in a six-dimensional embedding of handwritten "2"s (Fig. 1B) preserve continuity not only in the visual features of loop and arch articulation, but also in the implied pen trajectories, which are the true degrees of freedom underlying those appearances.
Nonlinear Dimensionality Reduction by Locally Linear Embedding

Sam T. Roweis¹ and Lawrence K. Saul²

Many areas of science depend on exploratory data analysis and visualization. The need to analyze large amounts of multivariate data raises the fundamental problem of dimensionality reduction: how to discover compact representations of high-dimensional data. Here, we introduce locally linear embedding (LLE), an unsupervised learning algorithm that computes low-dimensional, neighborhood-preserving embeddings of high-dimensional inputs. Unlike clustering methods for local dimensionality reduction, LLE maps its inputs into a single global coordinate system of lower dimensionality, and its optimizations do not involve local minima. By exploiting the local symmetries of linear reconstructions, LLE is able to learn the global structure of nonlinear manifolds, such as those generated by images of faces or documents of text.

How do we judge similarity? Our mental representations of the world are formed by processing large numbers of sensory inputs—including, for example, the pixel intensities of images, the power spectra of sounds, and the joint angles of articulated bodies. While complex stimuli of this form can be represented by points in a high-dimensional vector space, they typically have a much more compact description. Coherent structure in the world leads to strong correlations between inputs (such as between neighboring pixels in images), generating observations that lie on or close to a smooth low-dimensional manifold. To compare and classify such observations—in effect, to reason about the world—depends crucially on modeling the nonlinear geometry of these low-dimensional manifolds.

Scientists interested in exploratory analysis or visualization of multivariate data (1) face a similar problem in dimensionality reduction. The problem, as illustrated in Fig. 1, involves mapping high-dimensional inputs into a low-dimensional “description” space with as many coordinates as observed modes of variability. Previous approaches to this problem, based on multidimensional scaling (MDS) (2), have computed embeddings that attempt to preserve pairwise distances [or generalized disparities (3)] between data points; these distances are measured along straight lines or, in more sophisticated usages of MDS such as Isomap (4), along shortest paths confined to the manifold of observed inputs. Here, we take a different approach, called locally linear embedding (LLE), that eliminates the need to estimate pairwise distances between widely separated data points. Unlike previous methods, LLE recovers global nonlinear structure from locally linear fits.

The LLE algorithm, summarized in Fig. 2, is based on simple geometric intuitions. Suppose the data consist of N real-valued vectors \( \mathbf{x}_i \), each of dimensionality \( D \), sampled from some underlying manifold. Provided there is sufficient data (such that the manifold is well-sampled), we expect each data point and its neighbors to lie on or close to a locally linear patch of the manifold. We characterize the local geometry of these patches by linear coefficients that reconstruct each data point from its neighbors. Reconstruction errors are measured by the cost function

\[
\varepsilon(W) = \sum_i \left( \mathbf{x}_i - \sum_j \mathbf{w}_{ij} \mathbf{x}_j \right)^2
\]

which adds up the squared distances between all the data points and their reconstructions. The weights \( W_{ij} \) summarize the contribution of the \( j \)th data point to the \( i \)th reconstruction. To compute the weights \( W_{ij} \), we minimize the cost

1. *Gatsby Computational Neuroscience Unit, University College London, 17 Queen Square, London WC1N 3AR, UK.* 2. *AT&T Lab—Research, 180 Park Avenue, Florham Park, NJ 07932, USA.* E-mail: roweis@gatsby.ucl.ac.uk (S.T.R.); lsaul@research.att.com (L.K.S.)


Fig. 1. The problem of nonlinear dimensionality reduction, as illustrated (10) for three-dimensional data (B) sampled from a two-dimensional manifold (A). An unsupervised learning algorithm must discover the global internal coordinates of the manifold without signals that explicitly indicate how the data should be embedded in two dimensions. The color coding illustrates the neighborhood-preserving mapping discovered by LLE; black outlines in (B) and (C) show the neighborhood of a single point. Unlike LLE, projections of the data by principal component analysis (PCA) (28) or classical MDS (2) map faraway data points to nearby points in the plane, failing to identify the underlying structure of the manifold. Note that mixture models for local dimensionality reduction (29), which cluster the data and perform PCA within each cluster, do not address the problem considered here: namely, how to map high-dimensional data into a single global coordinate system of lower dimensionality.
The Isomap Algorithm and Topological Stability

Tenenbaum et al. (1) presented an algorithm, Isomap, for computing a quasi-isometric, low-dimensional embedding of a set of high-dimensional data points. Two issues need to be raised concerning this work. First, the basic approach presented by Tenenbaum et al. is not new, having been described in the context of flattening cortical surfaces using geodesic distances and multidimensional scaling (2, 3) and in later work that used Dijkstra’s algorithm to approximate geodesic distances (4, 5). These ideas generalize to arbitrary dimensionality if the connectivity and metric information of the manifold are correctly supplied.

Second, and more important, this approach is topologically unstable and can only be used after careful preprocessing of the data (6). In the application domain of cortical flattening, it is necessary to check manually for connectivity errors, so that points nearby in 3-space (for example, on opposite banks of a cortical sulcus) are not taken to be nearby in the cortical surface. If such care is taken, this method represents the preferred method for quasi-isometric cortical flattening.

What is new about the Isomap algorithm is how it defines the connectivity of each data point via its nearest Euclidean neighbors in the high-dimensional space. This step is vulnerable to short-circuit errors if the neighborhood is too large with respect to folds in the manifold on which the data points lie or if noise in the data moves the points slightly off the manifold. Even a single short-circuit error can alter many entries in the geodesic distance matrix, which in turn can lead to a drastically different (and incorrect) low-dimensional embedding.

We illustrate this failure in Fig. 1, using the MATLAB code published by Tenenbaum et al. along with their “Swiss roll” data, to which we have added a small amount of noise (7). Clearly, the algorithm is topologically unstable: Small errors in the data connectivity (topology) can lead to large errors in the solution. Choosing a very small neighborhood is not a satisfactory solution, as this can fragment the manifold into a large number of disconnected regions. Choosing the neighborhood “just right” requires a priori information about the global geometry of the high-dimensional data manifold (8), but, presumably, it is exactly in the absence of such information that one would need to use an algorithm to find “meaningful low-dimensional structures hidden in high-dimensional observations” (1). In summary, the basic idea of Isomap has long been known, and the new component introduced by Tenenbaum et al. provides an unreliable estimate of surface connectivity, which can lead to failure of the algorithm to perform as claimed.

Mukund Balasubramanian
Department of Cognitive and Neural Systems
Department of Electrical and Computer Systems
Boston University
Boston, MA 02215, USA

Responses and Notes

5. For a surface represented as a polyhedral mesh, exact geodesics can be computed in exponential time (2). However, geodesics between all pairs of vertices on the mesh can be approximated in O(N log N) time using Dijkstra’s algorithm to compute the shortest path along the edges of the polygons in the mesh (4). For meshes with O(10^7) vertices, exact geodesics can be computed in about 10 min on a 500-MHz Pentium III PC; for much larger meshes, faster approximation schemes are desirable.
7. We added zero-mean, normally distributed noise to the coordinates of each point, with the standard deviation of the noise chosen to be 2% of smallest dimension of the bounding box enclosing the data set (Fig. 1). Using smaller errors (for example, 1%) also led to failures, but we have not systematically studied the precise nature of algorithm failure, algorithm parameters, noise definition, and degree of failure of the solution.
8. Tenenbaum et al. provide proofs to show that if quantities such as the minimal radius of curvature of the manifold, the minimal branch separation, the volume of the manifold, and the density of the data points are known, then the neighborhood size can be bounded so that the algorithm performs correctly [note 18 of (1)]. In practice, however, these quantities will not be known in advance for a given set of high-dimensional data points. Furthermore, there is no prescription for how to compute these bounds in the presence of noise.

Fig. 1. (A) The “Swiss roll” data used by Tenenbaum et al. (1) to illustrate their algorithm (n = 1000). (B) The two-dimensional (2D) representation computed by the ε-Isomap variant of the Isomap algorithm, with ε = 5. Nearby points in the 2D embedding are also nearby points in the 3D manifold, as desired. (C) Data shown in A, with zero-mean normally distributed noise added to the coordinates of each point, where the standard deviation of the noise was chosen to be 2% of smallest dimension of the bounding box enclosing the data. (D) The Isomap (ε = 5) solution for the noisy data. There are gross “folds” in the embedding, and neither the metric nor the topological structure of the solution in (B) is preserved.

23 April 2001; accepted 20 November 2001

Response: Balasubramanian and Schwartz claim that “the basic idea” of our Isomap technique for nonlinear dimensionality reduction (1) has “long been known” in the context of flattening cortical surfaces. However, the problem of cortical flattening differs in crucial ways from our problem of finding low-dimensional structure in a cloud of high-dimensional data points. State-of-the-art procedures for cortical flattening (2, 3) take as input a triangulated mesh of fixed topology and dimensionality, which represents additional structure not available in the general problem we attempted to solve. We take as input only a collection of unorganized data points; the topology and dimensionality of the underlying manifold are unknown and
Fig. 1. (A) Our Swiss roll data set (1), shaded according to one intrinsic dimension of the underlying manifold. (B) The Swiss roll data plus Gaussian noise, with zero mean and standard deviation equal to approximately 7.5% of the separation between branches of the manifold (comparable to the noise level used by Balasubramanian and Schwartz). (C) An Isomap embedding of the noiseless data in (A) with neighborhood size $\epsilon = 5$ preserves both the intrinsic geometry and the topological order of the underlying manifold. (D) An Isomap embedding of the noisy data in (B) with an appropriately chosen neighborhood size ($\epsilon = 4.6$) also preserves both the intrinsic geometry and the topological order of the underlying manifold, just as in (C). (E) An appropriate neighborhood size for embedding the noiseless data into 2D Euclidean space can be determined based on a trade-off between two cost functions: the fraction of the variance in geodesic distance estimates not accounted for in the Euclidean embedding ($\epsilon$: black triangles), and the fraction of points not included in the largest connected component of the neighborhood graph, and thus not included in the Euclidean embedding (gray filled circles, with vertical scale three times that shown). Setting the neighborhood size ($x$ axis) too large (above $\epsilon = 6.2$) masks the intrinsic low-dimensional structure of the data set, leading to geodesic distance estimates that are poorly represented by a 2D Euclidean embedding and thus a much higher residual variance. Setting the neighborhood size too small ($\epsilon$ too small) leads many points from the largest connected component of the neighborhood graph, leading to the deletion of those points from the Isomap solution and increased residual variance for the remaining points. A stable range of neighborhood sizes between these two extremes leads to embeddings of the entire data set with near-zero residual variance. (F) The same analysis can be used to select a reasonable neighborhood size for the noisy data. A neighborhood size that works well for the noiseless data ($\epsilon = 5$) is too large here, yielding a residual variance of 0.25. However, a stable range of neighborhood sizes just slightly smaller ($\epsilon$ between 3.5 and 4.6) yields reasonable results, with residual variances all less than or equal to 0.01.

Joshua B. Tenenbaum*
Department of Psychology
Stanford University
Stanford, CA 94305—2130, USA
E-mail: jbt@psych.stanford.edu
T E C H N I C A L   C O M M E N T S

5. Balasubramanian and Schwartz add Gaussian noise with standard deviation equal to approximately 7.5% of the minimal separation between branches of the manifold as embedded in the 3D observation space. We characterize the magnitude of the noise with respect to the manifold’s minimal branch separation, rather than the minimal extent of the data’s bounding box, as do Balasubramanian and Schwartz, because for nonlinear manifolds, the former parameter influences the susceptibility to short-circuit errors (1, 10) while the latter does not.

6. We use a measure of residual variance [note 42 of (1)] to characterize how well the low-dimensional Euclidean embedding captures the geodesic distances estimated from the neighborhood graph. Lower residuals indicate better-fitting solutions, with less metric distortion.

7. It is possible to analyze some of the effects of noise through an extension of our asymptotic convergence arguments for noiseless data (1, 10). The principal danger with noisy data is that short-circuit edges may appear in the neighborhood graph, significantly changing its topology. In the noiseless situation, there is an upper bound $\epsilon_{\text{max}}$ for $\epsilon$, below which we avoid these gross topological errors. If the noise is bounded by a constant $c$, then we must reduce this upper bound to $\epsilon_{\text{max}} - 2c$. For Gaussian noise, one looks for a value of $c$ where the probability that the noise is not bounded by $c$ is as small as required. The calculation depends on the number of data points $N$ and the variance of the noise; if the variance is not too large, this works reasonably well for realistic values of $N$.


9. We have not studied in detail the behavior of LLE with noisy data, but in the cases we have explored, its behavior seems to be qualitatively similar to that of Isomap. For instance, a neighborhood size of $\epsilon = 5.5$ allows LLE to construct a topology-preserving embedding for the noiseless Swiss roll data in Fig. 1A but yields substantial topological errors for the noisy data in Fig. 1B.


11. Supported by NSF, the DARPA Human ID program, and ONR. We thank M. Bernstein and an anonymous referee for helpful comments.

30 August 2001; accepted 20 November 2001