ISOMAP for Dimensionality Reduction

A Global Geometric Framework for Nonlinear Dimensionality Reduction

Joshua B. Tenenbaum,[1]* Vin de Silva,[2] John C. Langford[3]

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 10^6 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.

Science

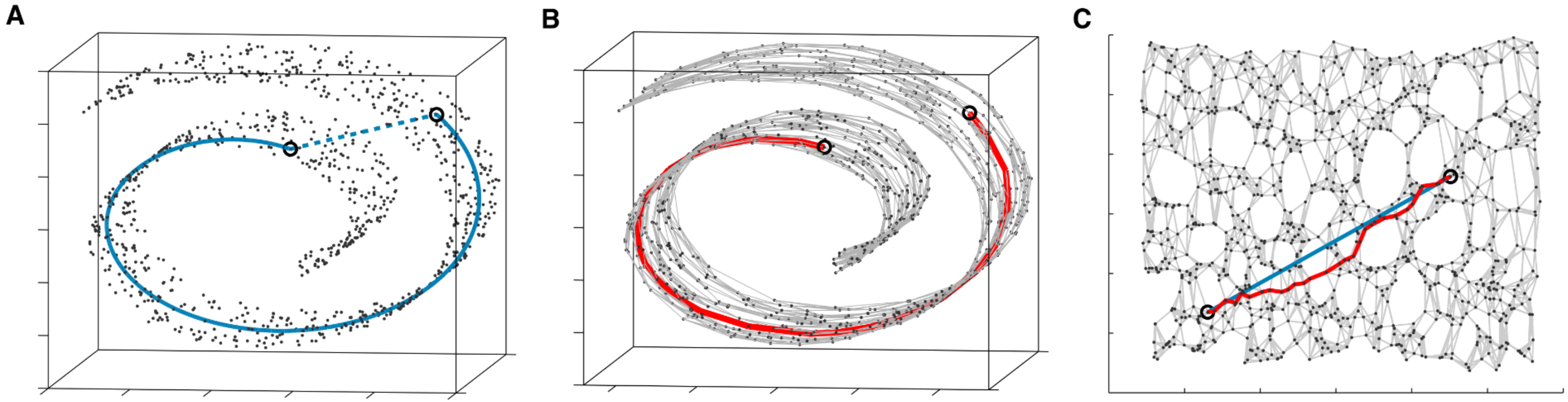
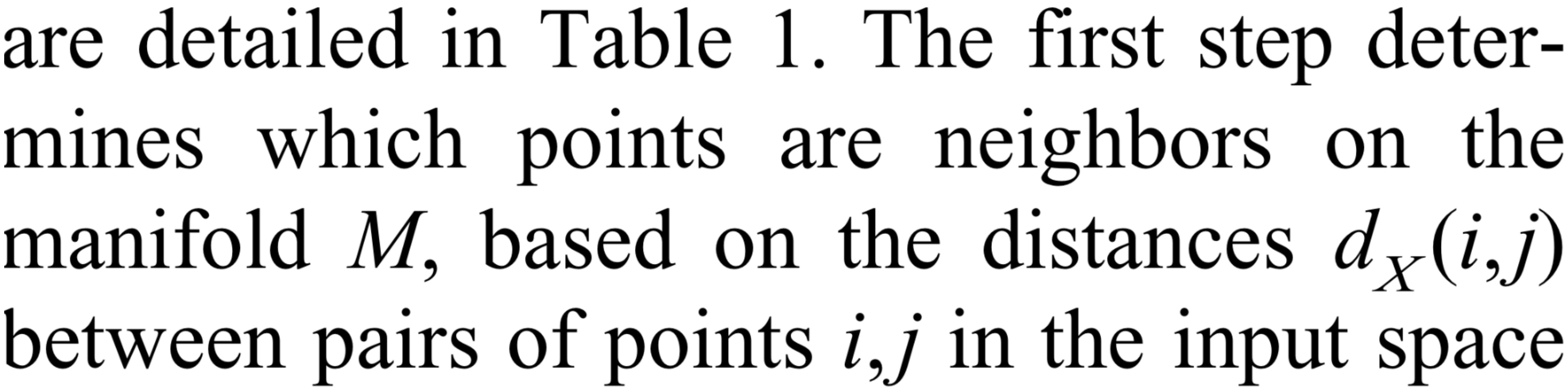


Fig. 3. The “Swiss roll” data set, illustrating how Isomap exploits geodesic paths for nonlinear dimensionality reduction. (**A**) For two arbitrary points (circled) on a nonlinear manifold, their Euclidean distance in the high-dimensional input space (length of dashed line) may not accurately reflect 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).

Three steps



are detailed in Table 1. The first step determines which points are neighbors on the manifold M, based on the distances $d_X(i,j)$ 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 ϵ, 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 $d_X(i,j)$ between neighboring points (Fig. 3B).

In its second step, Isomap estimates the geodesic distances $d_M(i,j)$ 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 $\mathbf{y}_i$ for points in Y are chosen to minimize the cost function

$$E = \|\boldsymbol{\tau}(D_G) - \boldsymbol{\tau}(D_Y)\|_{L^2} \qquad (1)$$

where D_Y denotes the matrix of Euclidean distances $\{d_Y(i,j) = \|\mathbf{y}_i - \mathbf{y}_j\|\}$ and $\|A\|_{L^2}$ the L^2 matrix norm $\sqrt{\Sigma_{i,j} A_{ij}^2}$. The $\boldsymbol{\tau}$ operator