Computational Data Analysis

Machine Learning

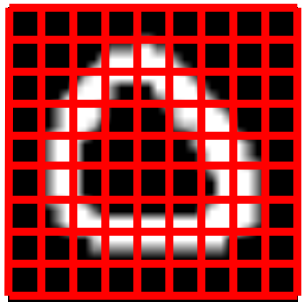
Yao Xie, Ph.D.
Associate Professor
Harold R. and Mary Anne Nash Early Career Professor
H. Milton Stewart School of Industrial and Systems Engineering

Nonlinear Dimensionality Reduction

What is dimensionality reduction?

The process of reducing the number of random variables under consideration

- One can combine, transform or select variables
- One can use linear or nonlinear operations



vector in R^n

Original data point

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$$f(x): R^n \mapsto R^k$$

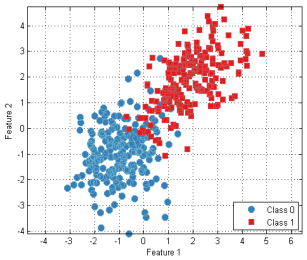
Reduced representation

$$z = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \end{pmatrix}$$

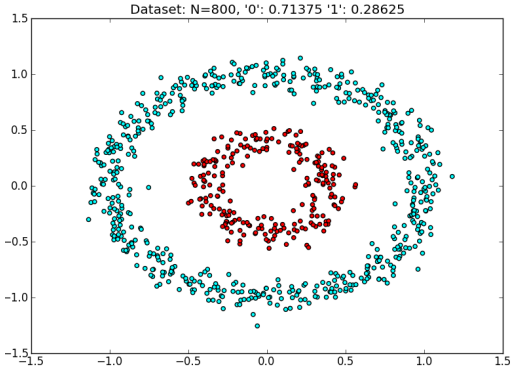
$$k \ll n$$

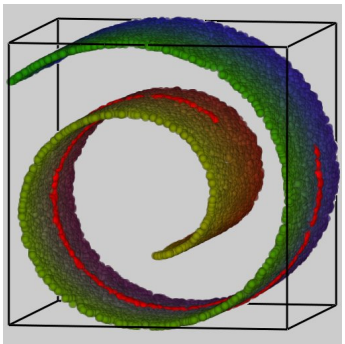
Limitation of PCA and SVD

- Suitable when variables are linearly correlated



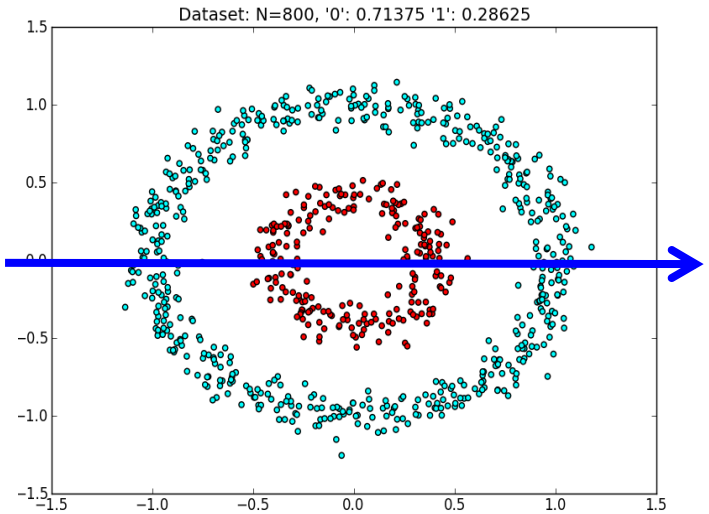
- Not suitable when nonlinear structures are present

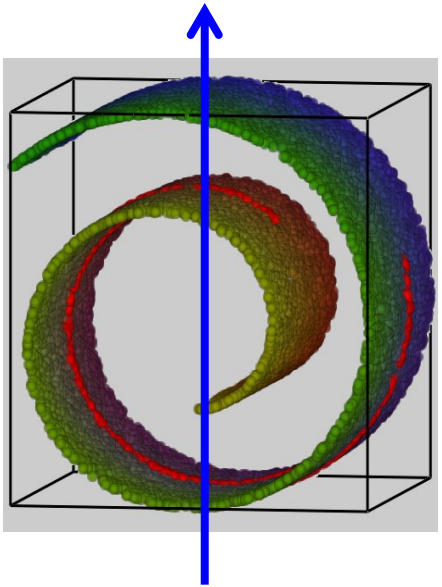




http://www.datawrangling.org/python-montage-code-for-displaying-arrays/

Limitation of PCA and SVD



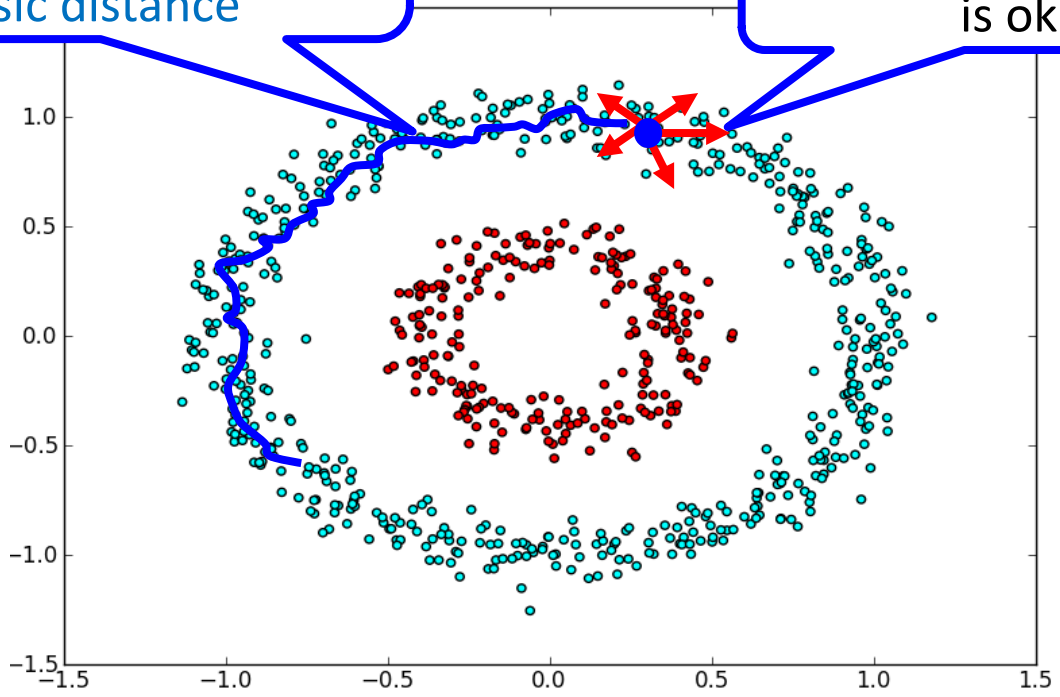


- PCA uses linear projection $w^{\top}x$, implicitly assuming Euclidean distance is the dissimilarity (distance) measure
- When there are nonlinear structure, Euclidean distance is not the right distance measure globally

What's a reasonable distance measure

long range distance like walking in data cloud (manifold)
Geodesic distance

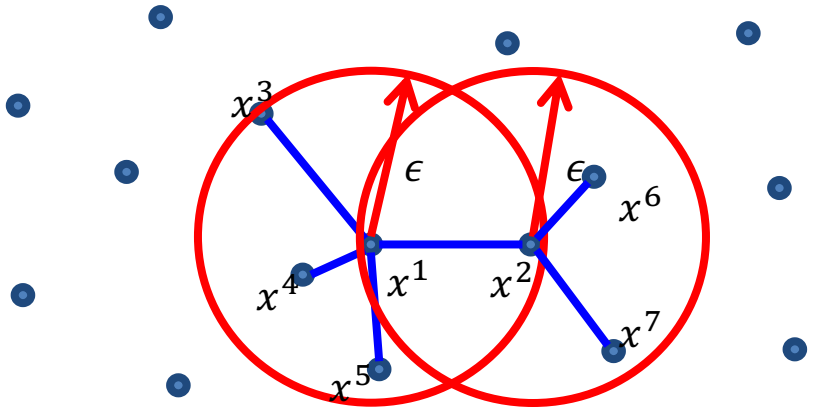
local Euclidean distance is ok



Weighted nearest neighbor graph

(ϵ-ISOMAP) Given m data points, threshold ϵ, construct matrix $A \in R^{m \times m}$

$$A^{ij} = \begin{cases} \left\| x^i - x^j \right\|, \text{if } \left\| x^i - x^j \right\| \leq \epsilon \\ 0, \qquad \text{otherwise} \end{cases}$$



(K-ISOMAP): select K nearest neighbors for each node

Isomap

Given m data points, $\{x^1, x^2, \ldots x^m\} \in R^n$, reduce data to k dimensional representation

Step 1: build a weighted graph A using nearest neighbors

Step 2: Compute pairwise shortest distance matrix D

entrywise square of the distance matrix

Step 3: use a centering matrix $H = I - \frac{1}{m} 11^\top$ to get

$$C = -\frac{1}{2} H (D)^2 H \in R^{m \times m} \qquad D_{ij}^2 := (D_{ij})^2$$

Step 4: compute leading eigenvectors $w^1, w^2, \ldots$ and eigenvalues $\lambda_1, \lambda_2 \ldots .$ of C

$$Z^T = (w^1, \ldots w^k) \begin{pmatrix} \lambda_1^{1/2} & & \\ & \ddots & \\ & & \lambda_k^{1/2} \end{pmatrix}$$

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.

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 thought of as points in a high-dimensional vector space, with each input dimension corresponding to the brightness of one pixel in the image or the firing rate of one retinal ganglion cell. Although the input dimensionality may be quite high (e.g., 4096 for these 64 pixel by 64 pixel images), the perceptually meaningful structure of these images has many fewer independent degrees of freedom. Within the 4096-dimensional input space, all of the images lie on an intrinsically three-dimensional manifold, or constraint surface, that can be parameterized by two pose variables plus an azimuthal lighting angle. Our goal is to discover, given only the unordered high-dimensional inputs, low-dimensional representations such as Fig. 1A with coordinates that capture the intrinsic degrees of freedom of a data set. This problem is of central importance not only in studies of vision (*1*–*5*), but also in speech (*6*, *7*), motor control (*8*, *9*), and a range of other physical and biological sciences (*10*–*12*).

[1]Department of Psychology and [2]Department of Mathematics, Stanford University, Stanford, CA 94305, USA. [3]Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15217, USA.

*To whom correspondence should be addressed. E-mail: jbt@psych.stanford.edu

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 (*13*). 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 (*4*, *5*, *11*, *14*). For example, both methods fail to detect the true degrees of freedom of the face data set (Fig. 1A), or even its intrinsic three-dimensionality (Fig. 2A).

Here we describe an approach that combines the major algorithmic features of 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-

ISOMAP key idea

Key idea: produce low dimensional representation which preserves "walking-distance" over the data cloud (manifold)

- Find neighbors $N(i)$ of each data point, x^i, within distance ϵ and let A be the adjacency matrix recording neighbor Euclidean distance
- Find shortest path distance matrix D between each pairs of points, x^i and x^j, based on A
- Find low dimensional representation which preserves the distances information in D

Idea: Manifold learning

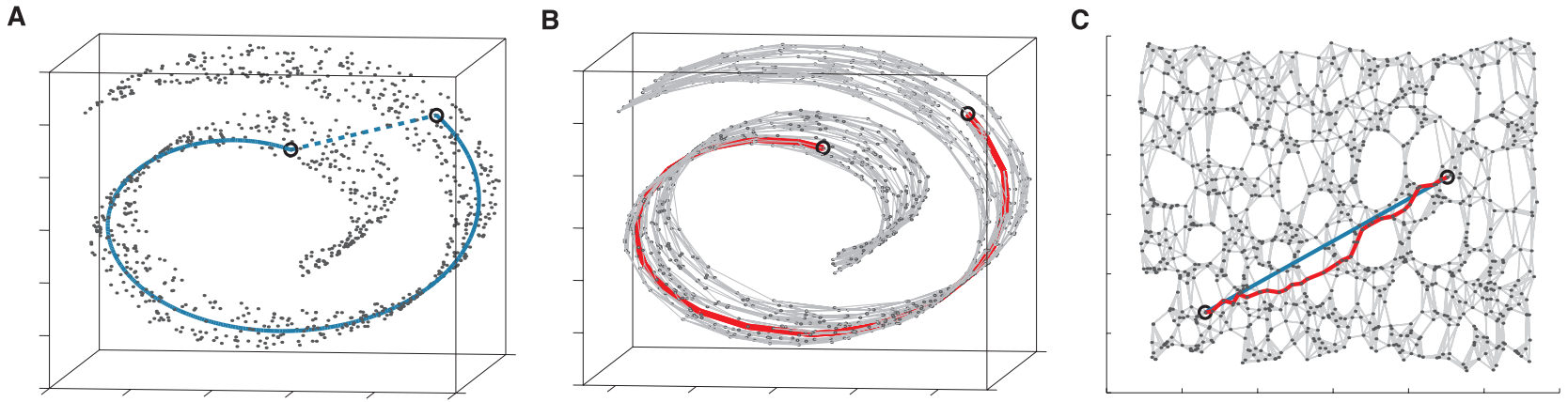


Fig. 3. The "Swiss roll" data set, illustrating how Isomap exploits geodesic paths for nonlinear dimensionality reduction. (**A**) For two arbitrary points (circled) on a nonlinear manifold, their Euclidean distance in the high-dimensional input space (length of dashed line) may not accurately reflect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

Shortest distance

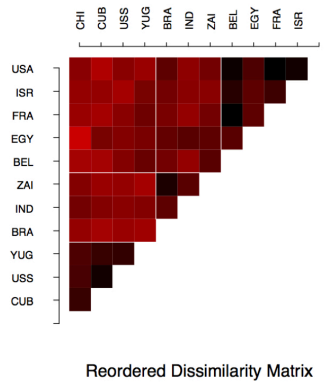
- With the graph defined by $A \in R^{m \times m}$
  - Find the shortest path distance matrix D between all pairs of points, also called graph distance matrix
- Can be computed with
  - Floyd-Warshall algorithm (all pair shortest path problem)
  - $m(m-1)/2$ applications of Dijkstra's algorithm

How to extract reduced representation

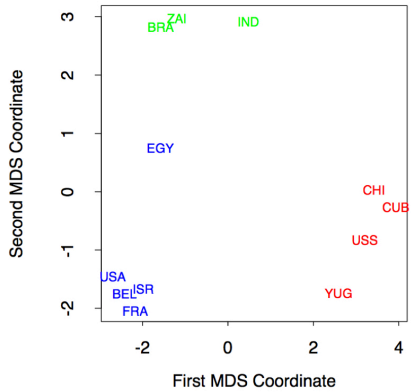
- Now we have a distance matrix D
- Now we can throw away the original data points and work with just D
- Can we extract from D a new set of coordinates (or reduced representation) Z which best explains D?
- Hints: related distance to inner product
  - Given coordinates, we can compute Euclidean distances
  - Given Euclidean distances, can we compute the coordinates?

MDS (Multi-dimensional scaling)

- Visualize/identify similarity pattern in data
- Goal of MDS: giving pairwise dissimilarity between data points, reconstruct a low-dimensional ”map” that preserves distances.
- From any dissimilarity (measures “dissimilar” two data points, no need to be induced by distance)
- Reconstructed map has coordinate $z^i = (z_1^i, \dots z_k^i)$ and natural Euclidean distance $||z^i - z^j||$



Reordered Dissimilarity Matrix



https://www.stat.pitt.edu/sungkyu/AAMA

MDS algorithm

- Distance to inner product

$$d_{ij}^2 = \left\|z^i - z^j\right\|^2 = \left(z^i - z^j\right)^T \left(z^i - z^j\right)$$
$$= \underbrace{z^{iT} z^i}_{} + \underbrace{z^{jT} z^j}_{} - \underbrace{2 z^{iT} z^j}_{}$$

Entries of Gram matrix $G = Z^T Z$

- Goal: Given a dissimilarity matrix $D = (d_{ij})$, MDS aims to find $z^1, \ldots, z^m \in R^k$ so that

$$d_{ij} \approx \left|\left|z^i - z^j\right|\right| \text{ as much as possible}$$

MDS algorithm

$$d_{ij}^2 = \left\|z^i - z^j\right\|^2 = \left(z^i - z^j\right)^T \left(z^i - z^j\right)$$
$$= {z^i}^T z^i + {z^j}^T z^j - 2{z^i}^T z^j$$

In matrix format, let $Z = (z^1, z^2, \ldots, z^m) \in R^{k \times m}$

$$(D)^2 = a1^\top + 1a^\top - 2Z^\top Z$$

where $a = \left({z^1}^\top z^1, {z^1}^\top z^1, \ldots, {z^m}^\top z^m\right)^\top$

Derivations (continued)

Construct a special centering matrix $H = I - \frac{1}{m}11^\top$

$$\left(I - \frac{1}{m}11^\top\right)1a^\top\left(I - \frac{1}{m}11^\top\right) = 0$$

$$\left(I - \frac{1}{m}11^\top\right)a1^\top\left(I - \frac{1}{m}11^\top\right) = 0$$

Then apply it to both side of $(D)^2$, it can be verified

$$-\frac{1}{2}H(D)^2H = -\frac{1}{2}H(a1^\top + 1a^\top - 2Z^\top Z)H = HZ^\top ZH = \tilde{Z}^T\tilde{Z}$$

$$\tilde{Z} \coloneqq ZH = Z\left(I - \frac{1}{m}11^\top\right) = Z - \frac{1}{m}\mu 1^\top \text{ (centered representation)}$$

Final step of ISOMAP

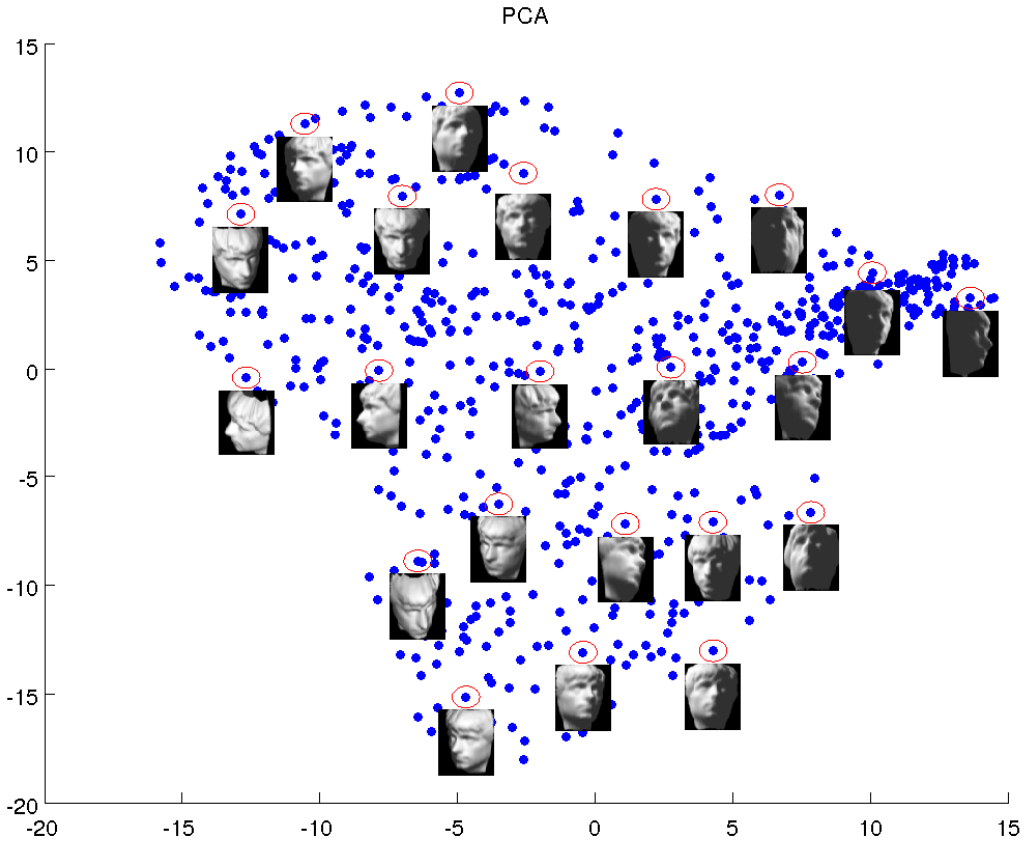
- Given $\tilde{G} := -\frac{1}{2}H(D)^2H \approx \tilde{Z}^\top\tilde{Z}$
- Perform eigendecomposition of

$$\tilde{G} = U\Lambda U^T$$

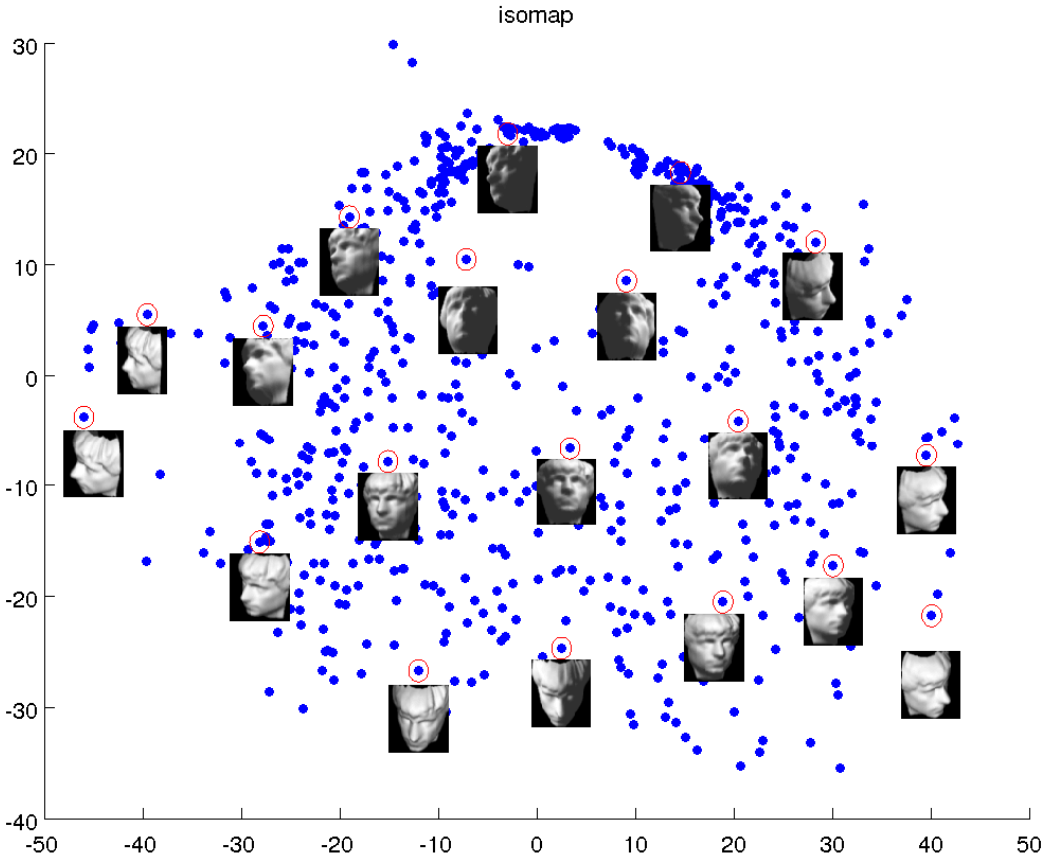
take the k (the dimension of the embedding space) leading eigenvalues and corresponding eigenvectors

- Reduced representation
  - $\tilde{Z}^\top = (u^1, \ldots, u^k)\begin{pmatrix} \lambda_1^{1/2} & & \\ & \ddots & \\ & & \lambda_k^{1/2} \end{pmatrix}$

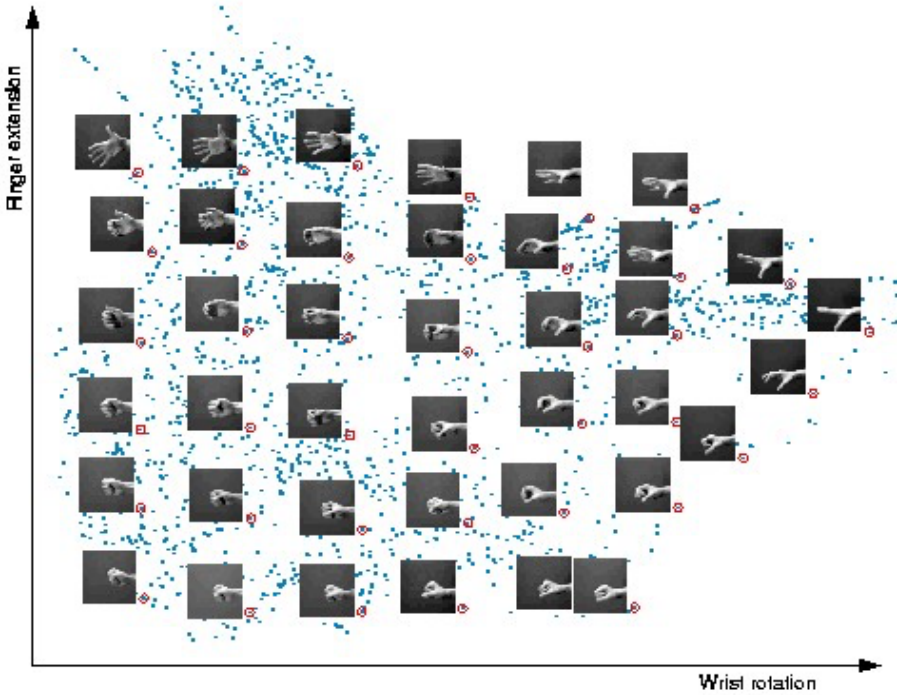
Is the principal direction interpretable?

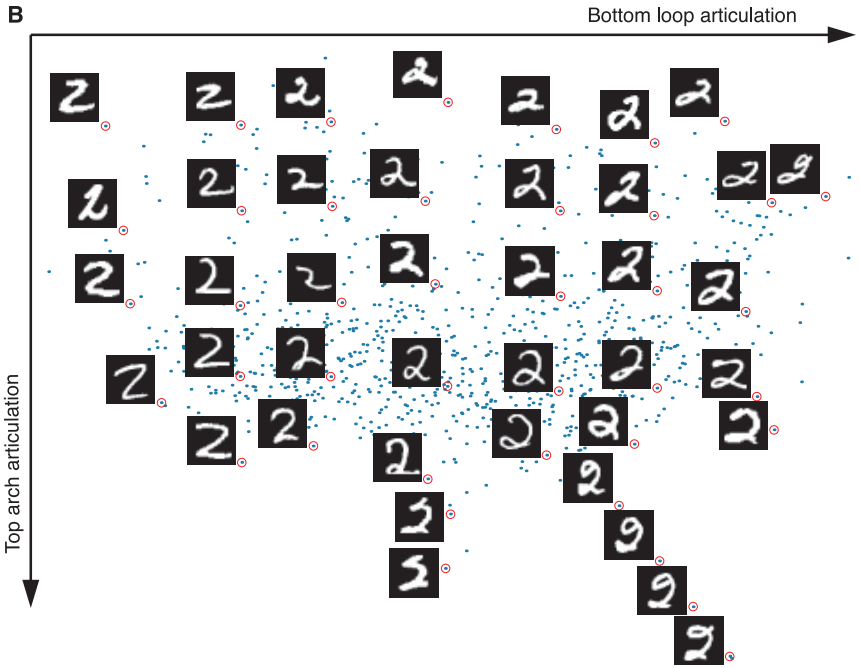


Result by isomap



More examples





https://web.mit.edu/cocosci/isomap/handfig.html

Other related topics

- Local linear embedding (LLE): low-dimensional, neighborhood preserving embedding of high-dimensional data (Roweis, Saul 2000)

$$\varepsilon(W) = \sum_{i} \left| \vec{X}_i - \Sigma_j W_{ij} \vec{X}_j \right|^2$$

- ISOMAP with spectral clustering for vectorial data: both using dissimilarity/similarity measures
- Kernel PCA (kPCA): ISOMAP and LLE can be viewed as kPCA with special kernels
- Manifold learning
- t-SNE

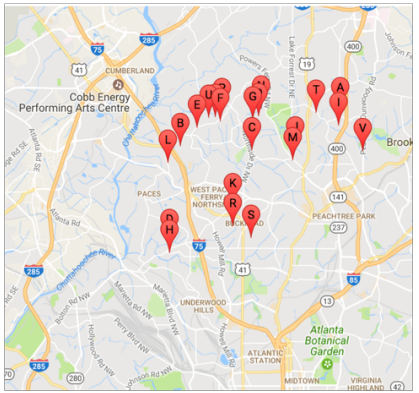
https://scikit-learn.org/stable/modules/manifold.html#:~:text=Manifold%20learning%20is%20an%20approach,sets%20is%20only%20artificially%20high.

Example: Atlanta police reports

Same dataset as last lecture
Text-embedding using RBM, visualization using **TSNE**

Id	Number	Category
Crime Series 1	8	*Robbery at Residence*
Crime Series 2	7	*Robbery at Gas Station*
Crime Series 3	4	*Pedestrian Robbery*
Crime Series 4	15	*Attempt Auto Theft*
Crime Series 5	22	*Burglary*
Random Cases	441	*Over 89 Categories*

22 cases of Buckhead burglary



https://arxiv.org/pdf/1902.00440


SUMMARY