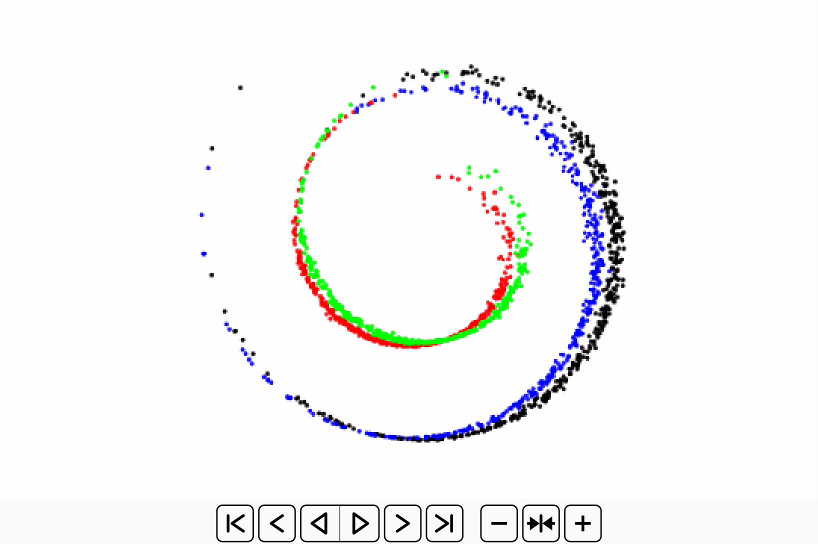
Manifold Learning

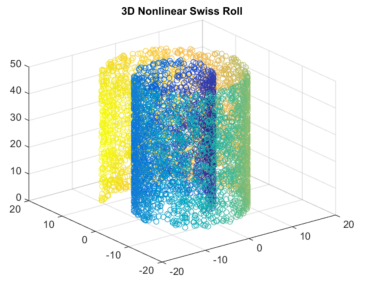
Spring 2018

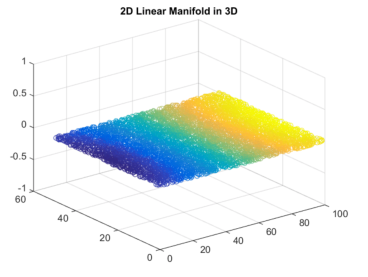
Weihang, Xu

Motivation

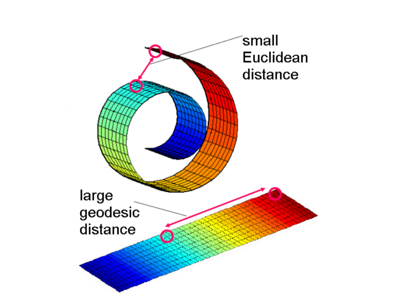


Swiss Roll Example





Swiss Roll Issues



- Consider finding clusters in this example.
- We could categorize data points based on their pairwise distance.
- However, these measurements in the original domain and may not approximate the geodesic distance on the manifold.

Manifold Learning

Manifold

Definition

Manifold in metric space : A manifold is a metric space M such that if $x \in M$, then there is some neighborhood U of x and some integer $n \geq 0$ such that U is homeomorphic to $\mathbb{R}^n$.

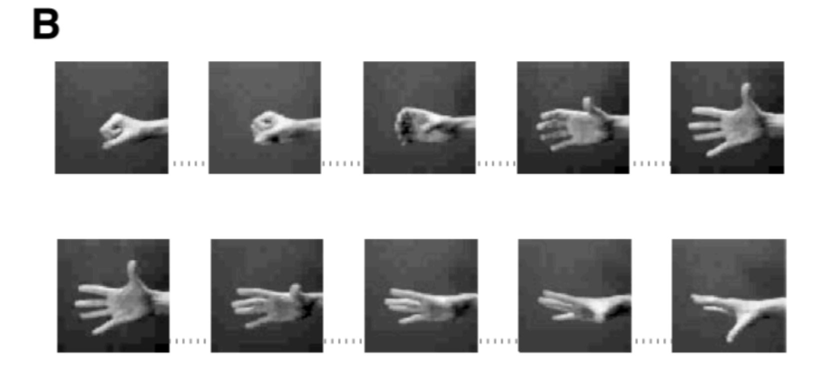
To speak English

- Manifolds are sets look locally like $\mathbb{R}^n$
- For instance, the unit circle in $\mathbb{R}^2$ is a manifold with dimension one.
- Why do we care about manifold?

An interesting case arises when the patterns of predictors lie on or near a low dimensional submanifold of the predictor space. In this case, the structure of the data set may be highly nonlinear, and linear methods are bound to fail.

Manifold Learning

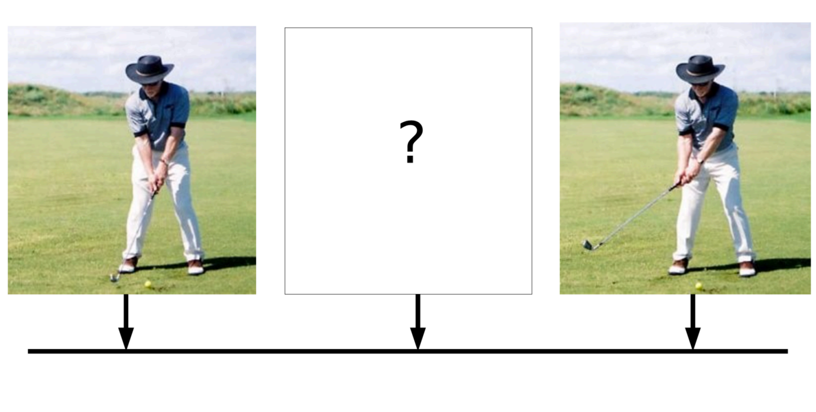
- Many real-world data sets are generated with very few degrees of freedom;
- An example is pictures of the same object under different imaging conditions, such as rotation angle or translation.



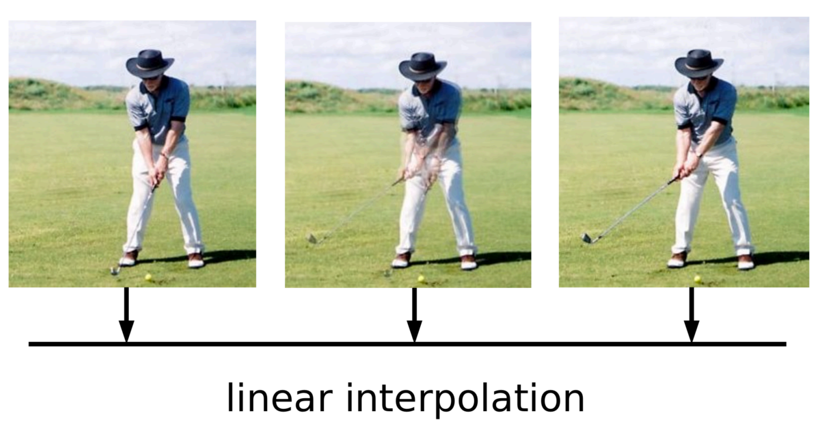
Manifold Learning

- When these factors vary smoothly, the data points can be seen as lying on a low-dimensional submanifold embedded in the high dimensional ambient space.
- Discovering such submanifolds and finding low-dimensional representations of them has been a focus of much recent work on unsupervised learning
- A key theme of these learning algorithms is to preserve (local) topological and geometrical properties (for example, geodesics, proximity, symmetry, angle) while projecting data points to low dimensional representations.

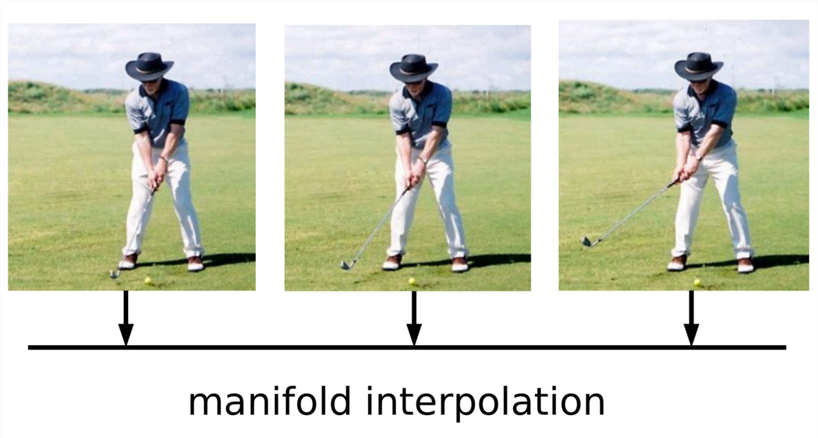
Example



Example



Example

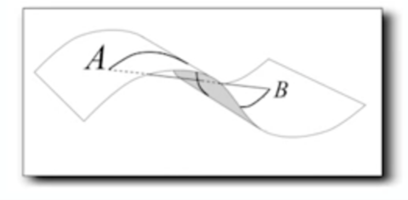


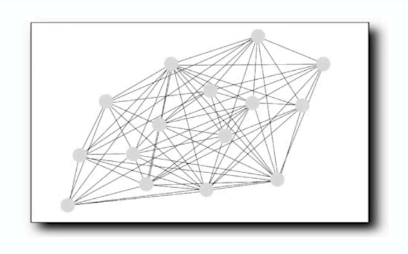
Isomap

Isomap

We want to approximate the geodesic distance on the manifold. Consider this problem with an undirected graph:

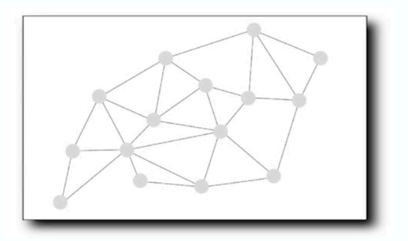
- Each vertex represent a data point
- Each edge corresponds to the pairwise distance of the two vertices.

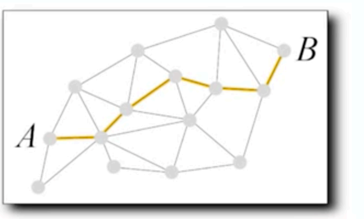




Isomap

- Choose a constraint: A threshold for the distance for each node.
- Remove edges from the graph, only keep those that satisfy the constraint.
- Use shortest path in graph to approximate the geodesic distance.



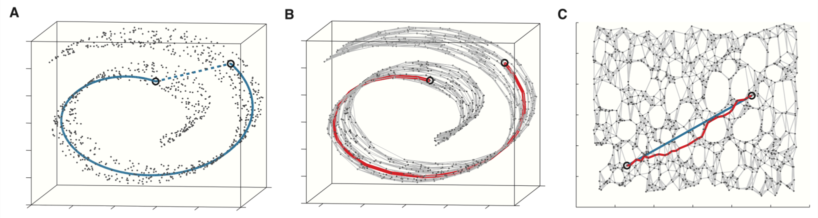


Isomap Algorithm

The algorithm has three steps:

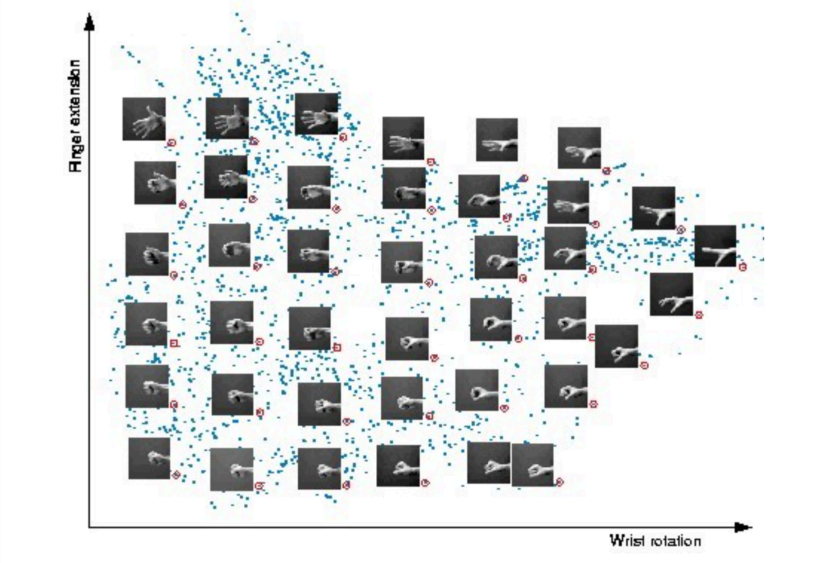
- Compute the k-nearest neighbors of each input pattern and to construct a graph whose vertices represent input patterns and whose (undirected) edges connect k-nearest neighbors. The edges are then assigned weights based on the Euclidean distance between nearest neighbors.
- Compute the pairwise distances $d_{ij} = \|x_i - x_j\|$ between all nodes (i, j) along shortest paths through the graph. (Djikstra's algorithm)
- The pairwise distances d_{ij} from Djikstra's algorithm are fed as input to MDS, yielding low dimensional outputs $z_i \in \mathbb{R}^m$ for which $\|z_i - z_j\| \approx d_{ij}$

Isomap Algorithm



When it succeeds, Isomap yields a low dimensional representation in which the Euclidean distances between outputs match the geodesic distances between input patterns on the submanifold from which they were sampled.

Isomap Algorithm



Isomap Pros and Cons

- preserves global structure
- few free parameters
- sensitive to noise, noise edges
- computationally expensive

Beyond Isomap

Multidimensional Scaling(MDS)

Multidimensional scaling seeks values $z_1, z_2, \cdots, z_n$ to minimize the so-called *stress function*.

$$S_M(z_1, z_2, \cdots, z_n) = \sum_{i \neq j} (d_{ij} - \|z_i - z_j\|)^2$$

This is known as *least squares* or *Kruskal–Shephard scaling*. The idea is to find a lower-dimensional representation of the data that preserves the pairwise distances as well as possible. Notice that the approximation is in terms of the distances rather than squared distances (which results in slightly messier algebra). A gradient descent algorithm is used to minimize S_M.

Maximum Variance Unfolding

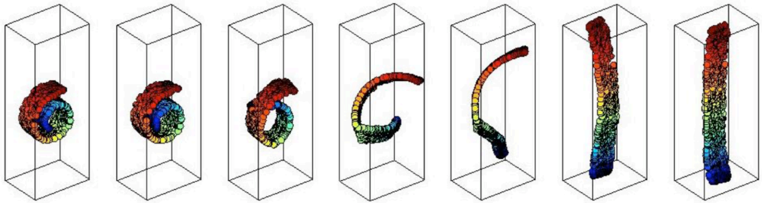


Figure 1.3 Input patterns sampled from a Swiss roll are "unfolded" by maximizing their variance subject to constraints that preserve local distances and angles. The middle snapshots show various feasible (but non-optimal) intermediate solutions of the optimization described in section 1.3.2.

Reference

Reference

Tenenbaum, Joshua B and De Silva, Vin and Langford, John C
A global geometric framework for nonlinear dimensionality reduction
science 290 5500 2319–2323, 2000.

Friedman, Jerome and Hastie, Trevor and Tibshirani, Robert.
The elements of statistical learning.
Springer series in statistics New York, 2001.

Spivak, Michael
A comprehensive introduction to differential geometry.
Publish or perish, 1979.

Saul, Lawrence K and Weinberger, Kilian Q and Ham, Jihun H and Sha, Fei and Lee, Daniel D