STAT 24620=FINM 34700, STAT 32950
Multivariate Data Analysis
Lecture 14: Nonlinear Unsupervised Learning

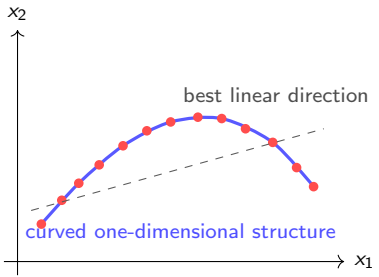
Jingshu Wang

The University of Chicago

Outline

When linear PCA is not enough

- PCA summarizes data using a low-dimensional **linear structure**.
- This works well when variation is approximately elliptical or planar.
- But some datasets have curved geometry:
  - images of a face under different rotations, where the images vary smoothly but not along a linear subspace of the pixel space;
  - points lying near a curved trajectory or spiral in a high-dimensional space.
- A linear projection may fail to follow the intrinsic curved structure of the data.



Can we generalize PCA by replacing a linear subspace with a smooth curve?

Principal curves: a nonlinear first component

- PCA summarizes data by a straight line:

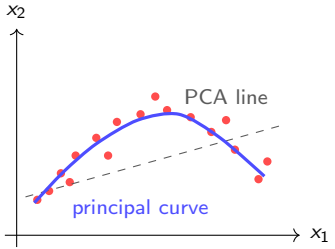
$$x_i \approx z_i v_1.$$

- A principal curve replaces the line by a smooth curve:

$$x_i \approx f(t_i), \qquad t_i \in \mathbb{R}.$$

- Think of t_i as a one-dimensional coordinate along the curve.
- Useful when the main variation follows a trajectory rather than a line.

Principal curves are a direct nonlinear analogue of the first principal component.



Fitting a principal curve

Let $f(t) \in \mathbb{R}^p$ be a smooth curve, and let t_i denote the location of x_i along the curve.

We would like to solve

$$\min_{f,\ t_1,\ldots,t_n} \sum_{i=1}^{n} \|x_i - f(t_i)\|^2 + \lambda\, \text{roughness}(f).$$

- Fit the data well
- Keep the curve smooth

Alternating fitting strategy

Initialize with PC1

$\Downarrow$

$t_i = \arg\min_t \|x_i - f(t)\|^2$ (project onto current curve)

$\Downarrow$

$x_{ij} \approx f_j(t_i), \quad j = 1, \ldots, p$ (fit f_j, smooth coordinates against t_i)

$\Downarrow$

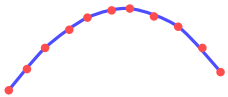
Update curve and iterate

Another perspective: geometry

Principal curves

$x_i \approx f(t_i)$

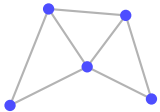
learn a smooth nonlinear structure directly



$\Longrightarrow$

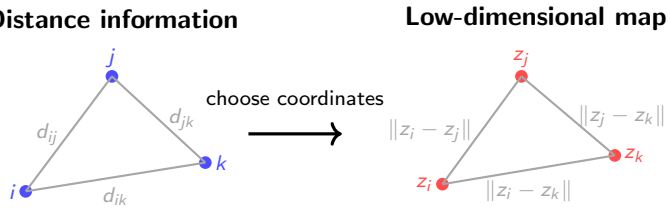
Geometry-based methods

preserve distances or neighborhoods



What does it mean for two points to be close?

MDS: preserving pairwise distances

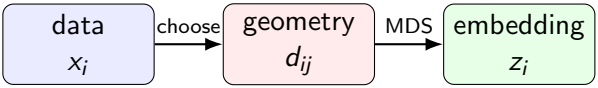


$$\|z_i - z_j\| \approx d_{ij}$$

- Classical MDS finds coordinates that preserve pairwise distances.
- With Euclidean distances, classical MDS gives essentially the same representation as PCA (up to rotation/reflection).
- Intuitively, PCA also preserves large-scale Euclidean geometry: distant points tend to remain distant after projection.
- Nonlinear methods often keep the embedding step but replace Euclidean distance by a better notion of distance.

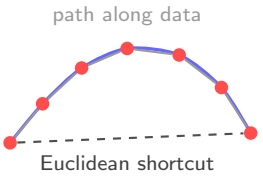
Changing the distance

- MDS separates the problem into two steps: choose distances, then embed.



- For curved structure, Euclidean distance can create a shortcut through empty space.
- A nonlinear route: replace Euclidean distance by distance along the data.

The key question: what distance captures the geometry of the data?



This motivates geodesic distance.

Isomap = MDS with geodesic distance

Isomap keeps the MDS embedding step, but changes the distance matrix.

1. Build a neighbor graph:

$$i \sim j \quad \text{if } x_j \text{ is among the } k \text{ nearest neighbors of } x_i.$$

2. Put edge weights equal to Euclidean distances between neighbors.
3. Approximate geodesic distances by shortest-path distances on the graph:

$$d_{ij}^{\text{geo}} = \text{shortest path distance from } i \text{ to } j.$$

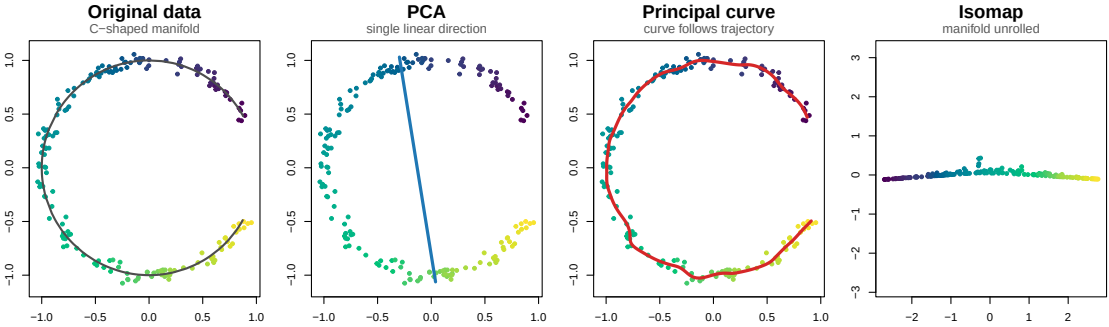
4. Apply classical MDS to D^{geo}.

Isomap = MDS using graph-based geodesic distances.

When does Isomap work well?

- Isomap works well when the data lie near a smooth low-dimensional manifold.
- The neighborhood graph should capture local geometry without creating shortcuts.
- Choosing the neighborhood size k is important:
  - too small: graph disconnects;
  - too large: curvature is lost.
- Isomap is less effective for disconnected clusters or highly noisy data.

Simulation comparison



- PCA gives the best linear direction, but it cannot follow the curved trajectory.
- A principal curve estimates the nonlinear trajectory directly.
- Isomap uses graph distances and tries to unroll the manifold in the embedding.

Summary

- PCA is linear; it can miss curved low-dimensional structure.
- Principal curves replace the first PC line by a smooth one-dimensional curve.
- MDS makes the role of pairwise distances explicit.
- Classical MDS with Euclidean distances gives a PCA-like embedding.
- Isomap replaces Euclidean distances with graph-based geodesic distances.
- The main tuning choice in Isomap is the neighborhood graph.

Changing the distance changes the geometry that the embedding tries to preserve.