Rehabilitating Isomap: Euclidean Representation of Geodesic Structure

Michael W. Trosset (mtrosset@indiana.edu) Gökçen Büyükbaş (gokbuyuk@indiana.edu)
Department of Statistics, Indiana University Department of Mathematics, Indiana University

View PDF

Isomap Procedure for Manifold Learning

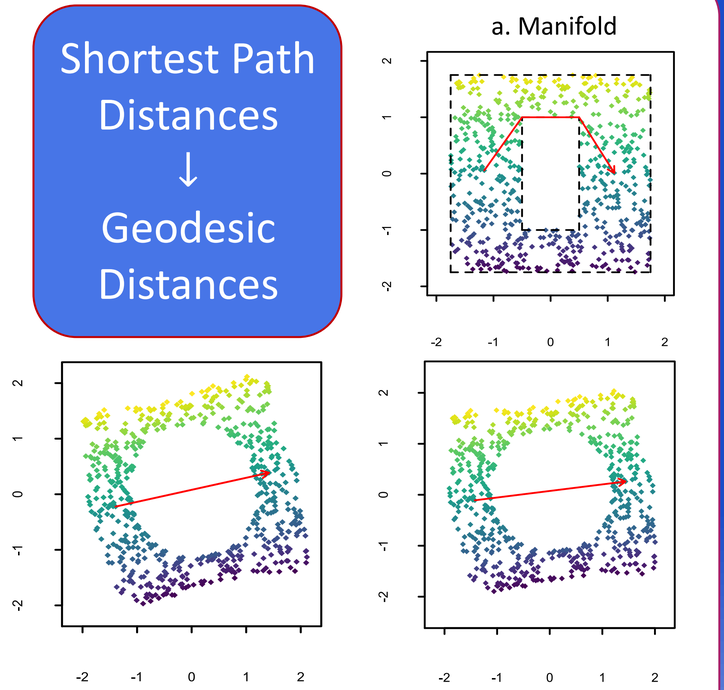
Given: feature vectors $x_1, \dots, x_n \in M \subset R^q$ and a target dimension d.

1. Construct an ϵ-neighborhood or K-nearest neighbor graph of the observed feature vectors. Weight the edge between $x_i \leftrightarrow x_j$ by $\|x_i - x_j\|$.
2. Compute the dissimilarity matrix $\Delta = [\delta_{jk}]$, where δ_{jk} is the shortest path distance on the graph between the vertices x_j and x_k. The key idea that underlies Isomap is that shortest path distances on a locally connected graph approximate Riemannian distances on the underlying Riemannian manifold M.
3. Embed Δ in R^d. Traditionally, Isomap embeds by classical multi-dimensional scaling (CMDS); however, if one's goal is to approximate shortest path distance with Euclidean distance, the one might prefer to embed differently, e.g., by minimizing Kruskal's raw stress criterion.

What is Manifold Learning?

"Just as PCA and MDS are guaranteed, given sufficient data, to recover the true structure of linear manifolds, Isomap is guaranteed asymptotically to recover the true dimensionality and geometric structure of a strictly larger class of nonlinear manifolds. Like the Swiss roll, these are manifolds whose intrinsic geometry is that of a convex region of Euclidean space, but whose ambient geometry in the high-dimensional input space may be highly folded, twisted, or curved. For non-Euclidean manifolds, such as a hemisphere or the surface of a doughnut, Isomap still produces a globally optimal low-dimensional Euclidean representation, as measured by Eq. 1." - (Tenenbaum et al., 2000)

Shortest Path Distances ↓ Geodesic Distances



b. Euclidean representation of Riemannian distances.

c. Euclidean representation of shortest path distances. (Isomap output with $\epsilon = 0.4$)

Is this a failure of Isomap?

- *"In the case of ISOMAP, the nonconvexity causes a strong dilation of the missing region, warping the rest of the embedding."* - (Donoho & Grimes, 2003)
- Or has Isomap constructed a reasonable Euclidean representation of the geodesic structure of a non-Euclidean manifold?

Manifold Learning as Parametrization Recovery

Suppose that $M = \psi(\Theta)$, where Θ is an open connected subset of R^p. The goal is to recover the original isometric coordinates θ, up to a rigid motion.

But the only manifolds that are locally isometric to any $\Theta \subset R^p$ have zero Gaussian curvature, i.e., Swiss Rolls. This is a very small subset of manifolds.

Perhaps Parametrization Recovery is not such a useful way to think about manifold learning?

Manifold Learning as Representing the Geodesic Structure

Theorem: Let $M \subset R^q$ be a compact connected Riemannian manifold and let μ be any probability measure on (M, B) such that $\mu(B(m,r)) > 0$ for every $m \in \mathrm{M}$ and $r > 0$. Suppose that $x_1, x_2, \dots \overset{iid}{\sim} \mu$ and let $V_n = \{x_1, \dots, x_n\}$.

- For $\epsilon > 0$, let $G_{n,\epsilon} = (V_n, E_{n,\epsilon})$ be a graph with vertex set V_n and edges between x_i&x_j if and only if $\|x_i - x_j\| \le \epsilon$.
- Let $d_{n,\epsilon}$ denote shortest path distance on $G_{n,\epsilon}$ with edge weights $\|x_i - x_j\|$.

Then there exist sequences $n_k \to \infty$ and $\epsilon_k \to 0$ for which

$$\sup_{m_a, m_b \in M} \inf_{x_a, x_b \in V_{n_k}} \left| d_{n_k,\epsilon_k}(x_a, x_b) - d_M(m_a, m_b) \right| \to^P 0$$

as $k \to \infty$.

Conclusion: Convexity is not required for Isomap to produce a useful low dimensional representation of the data.