

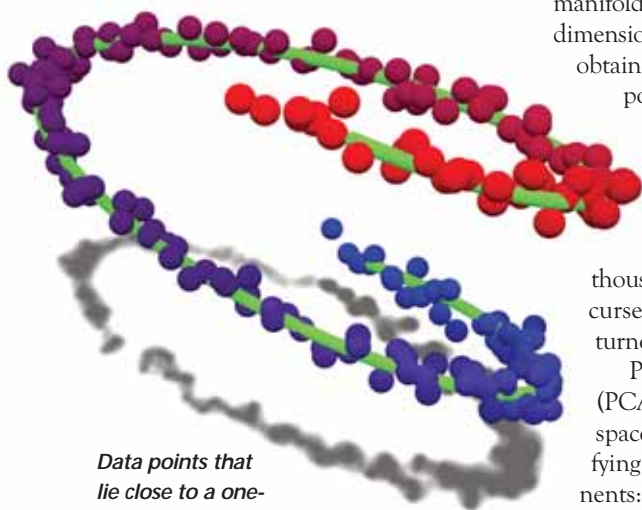
BY ZACHARY PINCUS

Dimension Reduction and Manifold Learning

When Less Is More

The Fall 2005 “Under the Hood” column discussed the *curse of dimensionality*—too many numerical components for each data point—and the *curse of dataset sparsity*—too few data points. One way to treat these problems in concert is to examine the geometric relationships between the data points, and represent the data with fewer descriptors that retain the salient structure.

This illustration demonstrates data in three dimensions that has such structure.



Data points that lie close to a one-dimensional manifold (green line).

DETAILS

Zach Pincus is a PhD Candidate in the Biomedical Informatics program at Stanford University. He works in the lab of Julie A. Theriot, developing methods for automatic and quantitative interpretation of images from microscopy, and cellular structures therein.

FOOTNOTE

¹Christopher Burge. “Geometric Methods for Feature Extraction and Dimensional Reduction: A Guided Tour,” in *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, Eds. L. Rokach and O. Maimon, Kluwer Academic Publishers.

ACKNOWLEDGEMENT: Thanks to Brian Naughton for many discussions of manifold learning.



Though each point is described with an (x,y,z) triplet, a single number—the parametric position along the spiral—may be sufficient to characterize that point for many applications. In technical terms, the points lie along a one-dimensional *manifold* (plus noise) that has been *embedded* in a three-dimensional space. Methods typically referred to as *manifold learning* or relating to *embeddings* seek to find such simpler parameterizations.

The simplest methods assume that the manifold is linear. Vast decreases in the dimensionality of the data points can be obtained by simply noting that n data points cannot span a linear space of dimension greater than $n-1$. For example, three points form no more than a two-dimensional plane, regardless of whether the points are vectors in ten or ten thousand dimensions. Thus, the two curses of biomedical data can be turned against one another.

Principal component analysis (PCA), for example, finds the subspace spanned by the data by identifying the data set’s principal components: The first such component is the direction along which the data has the most variance; the second component lies orthogonal to the first and best accounts for the remaining variation; and so on. The data in the illustration have three principal components: first, the central axis of the spiral; then, the long axis of the ellipse that remains when the spiral axis has been projected away; and finally the ellipse’s short axis. In many cases even highly nonlinear data can be described well

with a small subset of the leading principal components. For example, if the illustrated data were ten-dimensional but had the same spiral structure, three principal components would describe the data exactly. Alternatively, if the data were described only with the first one or two principal components, much meaningful structure would still be retained.

A related technique called multi-dimensional scaling (MDS) operates over the dis-

The two curses of biomedical data can be turned against one another.

tances between data points. MDS finds positions for the points in low dimensions such that the inter-point distances are changed as little as possible. If the Euclidian distances are provided, MDS and PCA are identical. However, other distance measures can also be constructed. The Isomap method finds the distance between points by measuring the length of a path that is constrained to “hop” from point to point along the data cloud. This distance approximates the *geodesic* distance (e.g., distance along the spiral). Application of MDS to these distances can easily recover many nonlinear structures.

In contrast to MDS and Isomap, which consider the distances from every point to every other, several methods, such as Laplacian eigenmaps and locally linear embedding (LLE), deal only with distances between points and their close neighbors.

It is my hope that this extremely fertile column has interested you in this fertile field of research. For a full-fledged introduction, refer to Christopher Burge’s excellent tutorial.¹ □