Nonlinear Dimensionality Reduction by Locally Linear Embedding Sam T. Roweis and Lawrence K. Saul Science 290 , 2323 (2000); DOI: 10.1126/science.290.5500.2323

*This copy is for your personal, non-commercial use only.*

If colleagues, you wish clients, to distribute or customers this article by

***clicking to others here. , you can order high-quality copies for your Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines here. The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 5, 2013 ): Updated version of information this article at:***

and services, including high-resolution figures, can be found in the online http://www.sciencemag.org/content/290/5500/2323.full.html

A list of selected additional articles on the Science Web sites related to this article can be found http://www.sciencemag.org/content/290/5500/2323.full.html#related

at:

This article cites 10 articles , 1 of which can be accessed free: http://www.sciencemag.org/content/290/5500/2323.full.html#ref-list-1

This article has been cited by 721 article(s) on the ISI Web of Science This article has been cited by 62 articles hosted by HighWire Press; see: http://www.sciencemag.org/content/290/5500/2323.full.html#related-urls

This Computers, article appears Mathematics

in the following subject collections: http://www.sciencemag.org/cgi/collection/comp\_math

Science American (print Association ISSN 0036-8075; for the Advancement online ISSN of 1095-9203) Science, 1200 is published New York weekly, Avenue except NW, Washington, the last week DC in December, 20005. Copyright by the 2000 by the American Association for the Advancement of Science; all rights reserved. The title Science is a registered trademark of AAAS.

R EPORTS

35. R. N. Shepard, Psychon. Bull. Rev. 1, 2 (1994). 36. J. B. Tenenbaum, Adv. Neural Info. Proc. Syst. 10, 682

(1998). 37. T. Martinetz, K. Schulten, Neural Netw. 7, 507 (1994).

1 es – in R2(D the ˆ M

, low-dimensional D

Y ). each algorithm. D Y

is the matrix of Euclidean distanc-

*Dˆ*

*M*

ing the coordinates of corresponding points {x

*i*

, y

*i*

www.sciencemag.org SCIENCE VOL290 22DECEMBER2000 2323

embedding recovered by

} in both spaces provided by Isomap together with stan- dard supervised learning techniques (39). 44. Supported by the Mitsubishi Electric Research Labo- 38. V. Kumar, A. Grama, A. Gupta, G. Karypis, Introduc-

ratories, the Schlumberger Foundation, the NSF tion to Parallel Computing: Design and Analysis of

(DBS-9021648), and the DARPA Human ID program. Algorithms (Benjamin/Cummings, Redwood City, CA,

We thank Y. LeCun for making available the MNIST 1994), pp. 257–297.

database and S. Roweis and L. Saul for sharing related 39. D. Beymer, T. Poggio, Science 272, 1905 (1996).

unpublished work. For many helpful discussions, we 40. Available at www.research.att.com/ yann/ocr/mnist.

thank G. Carlsson, H. Farid, W. Freeman, T. Griffiths, 41. P. Y. Simard, Y. LeCun, J. Denker, Adv. Neural Info.

R. Lehrer, S. Mahajan, D. Reich, W. Richards, J. M. Proc. Syst. 5, 50 (1993).

Tenenbaum, Y. Weiss, and especially M. Bernstein. 42. In order to evaluate the fits of PCA, MDS, and Isomap on comparable grounds, we use the residual variance

10 August 2000; accepted 21 November 2000

Nonlinear Dimensionality Reduction by Locally Linear Embedding

Sam T. Roweis1 and Lawrence K. Saul2

Many areas of science depend on exploratory data analysis and visualization. The need to analyze large amounts of multivariate data raises the fundamental problem of dimensionality reduction: how to discover compact representations of high-dimensional data. Here, we introduce locally linear embedding (LLE), an unsupervised learning algorithm that computes low-dimensional, neighbor- hood-preserving embeddings of high-dimensional inputs. Unlike clustering methods for local dimensionality reduction, LLE maps its inputs into a single global coordinate system of lower dimensionality, and its optimizations do not involve local minima. By exploiting the local symmetries of linear reconstruc- tions, LLE is able to learn the global structure of nonlinear manifolds, such as those generated by images of faces or documents of text.

How do we judge similarity? Our mental representations of the world are formed by processing large numbers of sensory in- puts—including, for example, the pixel in- tensities of images, the power spectra of sounds, and the joint angles of articulated bodies. While complex stimuli of this form can be represented by points in a high-dimensional vector space, they typically have a much more compact description. Coherent structure in the world leads to strong correlations between in- puts (such as between neighboring pixels in images), generating observations that lie on or close to a smooth low-dimensional manifold. To compare and classify such observations—in effect, to reason about the world—depends crucially on modeling the nonlinear geometry of these low-dimensional manifolds.

Scientists interested in exploratory analysis or visualization of multivariate data (1) face a similar problem in dimensionality reduction. The problem, as illustrated in Fig. 1, involves mapping high-dimensional inputs into a low- dimensional “description” space with as many

is each algorithm’s best estimate of the intrinsic manifold distances: for Isomap, this is the graph the Euclidean with the distance input-space matrix distance D

G ; for handwritten “2”s, where PCA and MDS, it is

matrix MDS D

X uses (except the tangent coefficient, distance). taken over R is the all entries standard of linear Dˆ

*M*

correlation and D

*Y*

. 43. In each sequence shown, the three intermediate im- ages are those closest to the points 1/4, 1/2, and 3/4 of the way between the given endpoints. We can also synthesize an explicit mapping from input space X to the low-dimensional embedding Y, or vice versa, us-

along shortest paths confined to the manifold of observed inputs. Here, we take a different ap- proach, called locally linear embedding (LLE), that eliminates the need to estimate pairwise distances between widely separated data points. Unlike previous methods, LLE recovers global nonlinear structure from locally linear fits.

The LLE algorithm, summarized in Fig. 2, is based on simple geometric intuitions. Suppose the data consist of N real-valued vectors pled from X i , some each of dimensionality D, underlying manifold. sam- Pro- vided there is sufficient data (such that the manifold is well-sampled), we expect each data point and its neighbors to lie on or close to a locally linear patch of the mani- fold. We characterize the local geometry of these patches by linear coefficients that reconstruct each data point from its neigh- bors. Reconstruction errors are measured by the cost function coordinates as observed modes of variability. Previous approaches to this problem, based on

ε W multidimensional scaling (MDS) (2), have

i computed embeddings that attempt to preserve pairwise distances [or generalized disparities (3)] between data points; these distances are measured along straight lines or, in more so- phisticated usages of MDS such as Isomap (4),

2

(1)

which adds up the squared distances between all the data points and their reconstructions. The weights W

ij

*X*

i j

*W*

ij

*X*

j

summarize the contribution of the jth data point to the ith reconstruction. To com- pute the weights W

ij

, we minimize the cost

1Gatsby Computational Neuroscience Unit, Universi- ty College London, 17 Queen Square, London WC1N 3AR, UK. 2AT&T Lab—Research, 180 Park Avenue, Florham Park, NJ 07932, USA.

E-mail: roweis@gatsby.ucl.ac.uk (S.T.R.); lsaul@research. att.com (L.K.S.)

Fig. 1. The problem of nonlinear dimensionality reduction, as illustrated (10) for three-dimensional data (B) sampled from a two-dimensional manifold (A). An unsupervised learning algorithm must discover the global internal coordinates of the manifold without signals that explicitly indicate how the data should be embedded in two dimensions. The color coding illustrates the neighborhood- preserving mapping discovered by LLE; black outlines in (B) and (C) show the neighborhood of a single point. Unlike LLE, projections of the data by principal component analysis (PCA) (28) or classical MDS (2) map faraway data points to nearby points in the plane, failing to identify the underlying structure of the manifold. Note that mixture models for local dimensionality reduction (29), which cluster the data and perform PCA within each cluster, do not address the problem considered here: namely, how to map high-dimensional data into a single global coordinate system of lower dimensionality.

R EPORTS

function subject to two constraints: first, that each data its neighbors point (5), X i enforcing is reconstructed W

ij

only from 0 if X

j

does

Fig. 2. Steps of locally lin- ear embedding: (1) Assign neighbors to each data point using the X

i

(for example by K nearest neigh- bors). (2) Compute the weights early its reconstruct W

ij neighbors, that best solving X

i

2324

22DECEMBER2000 VOL290 SCIENCE www.sciencemag.org lin- from the constrained least-squares problem in Eq. 1. (3) Com- pute the low-dimensional embedding reconstructed mizing Eq. 2 vectors by by finding W

ij

Y , i

mini- best

the smallest eigenmodes of the sparse symmetric ma- trix in Eq. 3. Although the weights are computed W

ij

and by vectors methods Y

i

in linear algebra, the con- straint that points are only reconstructed from neigh- bors can result in highly nonlinear embeddings.

Fig. 3. Images of faces (11) mapped into the embedding space described by the first two coordinates of LLE. Representative faces are shown next to circled points in different parts of the space. The bottom images correspond to points along the top-right path (linked by solid line), illustrating one particular mode of variability in pose and expression.

; not belong to the set of neighbors of second, that the rows of the weight matrix

*X*

i

sum to one:

j

*W*

ij

1. The optimal weights

subject W by ij

solving a to least-squares these constraints problem (6) are (7).

found

The constrained weights that minimize these reconstruction errors obey an important symmetry: for any particular data point, they are invariant to rotations, rescalings, and translations of that data point and its neigh- bors. By symmetry, it follows that the recon- struction weights characterize intrinsic geo- metric properties of each neighborhood, as opposed to properties that depend on a par- ticular frame of reference (8). Note that the invariance to translations is specifically en- forced by the sum-to-one constraint on the rows of the weight matrix.

Suppose the data lie on or near a smooth nonlinear manifold of lower dimensionality d D. To a good approximation then, there exists a linear mapping—consisting of a translation, rotation, and rescaling—that maps the high-dimensional coordinates of each neighborhood to global internal coordi- nates on the manifold. By design, the recon- struction ric properties weights of the W

ij

data reflect that intrinsic are invariant geomet- to exactly such transformations. We therefore expect their characterization of local geome- try in the original data space to be equally valid for local patches on the manifold. In particular, struct the same weights the ith data point in W

D ij

dimensions that recon-

should also reconstruct its embedded mani- fold coordinates in d dimensions.

LLE constructs a neighborhood-preserving mapping based on the above idea. In the final step of the algorithm, each high-dimensional observation vector nates on Y

i the representing X manifold. i

is mapped to a low-dimensional global internal coordi- This is done by choosing d-dimensional coordinates embedding cost function

*Y*

i

to minimize the

*Y*

i

2 Y

i j

*W*

ij

*Y*

j

(2)

This cost function, like the previous one, is based on locally linear reconstruction errors, but mizing here we fix the weights W

ij

while opti- cost in the coordinates Eq. 2 defines a quadratic Y

i

. The embedding form in the vectors the problem Y

i

. Subject to constraints that make well-posed, it can be minimized by solving a sparse N N eigenvalue prob- lem (9), whose bottom d nonzero eigenvec- tors provide an ordered set of orthogonal coordinates centered on the origin.

Implementation of the algorithm is straightforward. In our experiments, data points were reconstructed from their K near- est neighbors, as measured by Euclidean dis- tance or normalized dot products. For such implementations of LLE, the algorithm has only one free parameter: the number of neighbors, K. Once neighbors are chosen, the optimal weights W

ij

and coordinates Y

i

are

computed by standard methods in linear al- gebra. The algorithm involves a single pass through the three steps in Fig. 2 and finds global minima of the reconstruction and em- bedding costs in Eqs. 1 and 2.

In addition to the example in Fig. 1, for which the true manifold structure was known (10), we also applied LLE to images of faces (11) and vectors of word-document counts (12). Two-dimensional embeddings of faces and words are shown in Figs. 3 and 4. Note how the coordinates of these embedding spaces are related to meaningful attributes, such as the pose and expression of human faces and the semantic associations of words. Many popular learning algorithms for nonlinear dimensionality reduction do not share the favorable properties of LLE. Itera- tive hill-climbing methods for autoencoder neural networks (13, 14), self-organizing maps (15), and latent variable models (16) do not have the same guarantees of global opti- mality or convergence; they also tend to in- volve many more free parameters, such as learning rates, convergence criteria, and ar-

Fig. 4. Arranging words in a continuous semantic space. Each word was initially repre- sented by a high-dimensional vector that counted the number of times it appeared in different encyclopedia ar- ticles. LLE was applied to these word-document count vectors (12), resulting in an embedding location for each word. Shown are words from two different bounded re- gions (A) and (B) of the em- bedding space discovered by LLE. Each panel shows a two- dimensional projection onto the third and fourth coordi- nates of LLE; in these two dimensions, the regions (A) and (B) are highly over- lapped. The inset in (A) shows a three-dimensional projection onto the third, fourth, and fifth coordinates, revealing an extra dimension along which regions (A) and (B) are more separated. Words that lie in the inter- section of both regions are capitalized. Note how LLE co- locates words with similar contexts in this continuous semantic space.

R EPORTS

chitectural specifications. Finally, whereas

alyzed—can provide information about global other nonlinear methods rely on deterministic

geometry. Many virtues of LLE are shared by annealing schemes (17) to avoid local mini-

Tenenbaum’s algorithm, Isomap, which has ma, the optimizations of LLE are especially

been successfully applied to similar problems in tractable.

nonlinear dimensionality reduction. Isomap’s LLE scales well with the intrinsic mani-

embeddings, however, are optimized to pre- fold dimensionality, d, and does not require a

serve geodesic distances between general pairs discretized gridding of the embedding space.

of data points, which can only be estimated by As more dimensions are added to the embed-

computing shortest paths through large sublat- ding space, the existing ones do not change,

tices of data. LLE takes a different approach, so that LLE does not have to be rerun to

analyzing local symmetries, linear coefficients, compute higher dimensional embeddings.

and reconstruction errors instead of global con- Unlike methods such as principal curves and

straints, pairwise distances, and stress func- surfaces (18) or additive component models

tions. It thus avoids the need to solve large (19), LLE is not limited in practice to mani-

dynamic programming problems, and it also folds of extremely low dimensionality or

tends to accumulate very sparse matrices, codimensionality. Also, the intrinsic value of

whose structure can be exploited for savings in d can itself be estimated by analyzing a re-

time and space. ciprocal cost function, in which reconstruc-

LLE is likely to be even more useful in tion weights derived from the embedding

combination with other methods in data anal- vectors LLE Y

illustrates i

are applied to a general the data principle points of mani-

*X*

i

.

ysis and statistical learning. For example, a parametric mapping between the observation fold learning, elucidated by Martinetz and

and embedding spaces could be learned by Schulten (20) and Tenenbaum (4), that over-

supervised neural networks (21) whose target lapping local neighborhoods—collectively an-

values are generated by LLE. LLE can also be generalized to harder settings, such as the case of disjoint data manifolds (22), and spe- cialized to simpler ones, such as the case of time-ordered observations (23).

Perhaps the greatest potential lies in ap- plying LLE to diverse problems beyond those considered here. Given the broad appeal of traditional methods, such as PCA and MDS, the algorithm should find widespread use in many areas of science.

References and Notes 1. M. L. Littman, D. F. Swayne, N. Dean, A. Buja, in Computing Science and Statistics: Proceedings of the 24th Symposium on the Interface, H. J. N. Newton, Ed. (Interface Foundation of North America, Fairfax Sta- tion, VA, 1992), pp. 208–217. 2. T. Cox, M. Cox, Multidimensional Scaling (Chapman &

Hall, London, 1994). 3. Y. Takane, F. W. Young, Psychometrika 42, 7 (1977). 4. J. Tenenbaum, in Advances in Neural Information Processing 10, M. Jordan, M. Kearns, S. Solla, Eds. (MIT Press, Cambridge, MA, 1998), pp. 682–688. 5. The set of neighbors for each data point can be assigned in a variety of ways: by choosing the K nearest neighbors in Euclidean distance, by consider- ing all data points within a ball of fixed radius, or by using prior knowledge. Note that for fixed number of neighbors, the maximum number of embedding di- mensions LLE can be expected to recover is strictly less than the number of neighbors. 6. For certain applications, one might also constrain the weights to be positive, thus requiring the reconstruc- tion of each data point to lie within the convex hull of its neighbors. 7. Fits: The constrained weights that best reconstruct each data point from its neighbors can be computed in closed form. Consider a particular data point x with neighbors w minimized j

. The reconstruction j

and sum-to-one reconstruction error  x – in three steps. First, evaluate weights

K j 1

inner w

j j

prod- 2 is

ucts between neighbors to compute the neighbor- hood correlation inverse, C 1. Second, matrix, compute C

jk the j k

Lagrange , and its matrix multi- plier, / , that enforces the sum-to-one con- straint, where 1

jk

) and

jk

www.sciencemag.org SCIENCE VOL290 22DECEMBER2000 2325

1( 1. C

jk

*x*

k

*w*

j C

jk

k

Third, compute the reconstruction weights: C

jk

1( x

k

). If the correlation matrix C is

R EPORTS

22DECEMBER2000 VOL290 SCIENCE www.sciencemag.org 2326

nearly singular, it can be conditioned (before inver- sion) by adding a small multiple of the identity matrix. This amounts to penalizing large weights that exploit correlations beyond some level of precision in the data sampling process. 8. Indeed, LLE does not require the original data to be described in a single coordinate system, only that each data point be located in relation to its neighbors. 9. The embedding vectors Y

i

are found by minimizing the cost function weights constraints W

ij that . This (Y)

optimization i

 Y

i

– make the problem j

W is ij

performed Y

j

2 over Y

i subject with fixed to well posed. It is clear that the displacement coordinates without Y i affecting can be translated the by a constant cost, (Y). We re- move this degree of freedom by requiring the coordi- nates to be centered on avoid degenerate solutions, the we origin:

constrain i

*Y*

i

0. Also, to the embed- ding ucts d vectors that identity to satisfy

matrix. have 1

*N*

unit Now Y i

covariance, with outer prod- R the Y i cost I, defines where I is the d

a quadratic form, (Y)

ij

i

the embedding vectors M

ij

( Y

i

*Y*

j ), and involving the inner products symmetric N of N matrix

*M*

ij ij

*W*

ij

*W*

ji

k

*W*

ki

*W*

kj

(3)

where embedding, ij

is 1 if i j and 0 otherwise. The optimal up to a global rotation of the embedding space, is found by computing the bottom d 1 eigenvectors of this matrix (24). The bottom eigen- vector of this matrix, which we discard, is the unit vector with all equal components; it represents a free translation mode of eigenvalue zero. (Discarding it enforces the constraint that the embeddings have zero mean.) The remaining d eigenvectors form the d embedding coordinates found by LLE. Note that the matrix M can be stored and manipulated as the sparse matrix (I W)T(I W), giving substantial computational savings for large values of N. More- over, its bottom d 1 eigenvectors (those corre- sponding to its smallest d 1 eigenvalues) can be found efficiently without performing a full matrix diagonalization (25).

10. Manifold: Data points in Fig. 1B (N 2000) were sampled from the manifold (D 3) shown in Fig. 1A. Nearest neighbors (K 20) were determined by Euclidean distance. This particular manifold was in- troduced by Tenenbaum (4), who showed that its global structure could be learned by the Isomap algorithm. 11. Faces: Multiple photographs (N 2000) of the same face were digitized as 20 28 grayscale images. Each image was treated by LLE as a data vector with D 560 elements corresponding to raw pixel intensities. Nearest neighbors (K 12) were determined by Euclidean distance in pixel space. 12. Words: Word-document counts were tabulated for N 5000 words from D 31,000 articles in Grolier’s Encyclopedia (26). Nearest neighbors (K 20) were determined by dot products between count vectors normalized to unit length. 13. D. DeMers, G. W. Cottrell, in Advances in Neural Information Processing Systems 5, D. Hanson, J. Cowan, L. Giles, Eds. (Kaufmann, San Mateo, CA, 1993), pp. 580–587. 14. M. Kramer, AIChE J. 37, 233 (1991). 15. T. Kohonen, Self-Organization and Associative Mem-

ory (Springer-Verlag, Berlin, 1988). 16. C. Bishop, M. Svensen, C. Williams, Neural Comput.

10, 215 (1998). 17. H. Klock, J. Buhmann, Pattern Recognition 33, 651

(1999). 18. T. J. Hastie, W. Stuetzle, J. Am. Stat. Assoc. 84, 502

(1989). 19. D. J. Donnell, A. Buja, W. Stuetzle, Ann. Stat. 22, 1635

(1994). 20. T. Martinetz, K. Schulten, Neural Networks 7, 507

(1994). 21. D. Beymer, T. Poggio, Science 272, 1905 (1996). 22. Although in all the examples considered here, the data had a single connected component, it is possible to formulate LLE for data that lies on several disjoint manifolds, possibly of different underlying dimen- sionality. Suppose we form a graph by connecting each data point to its neighbors. The number of connected components (27) can be detected by ex-

amining powers of its adjacency matrix. Different connected components of the data are essentially decoupled in the eigenvector problem for LLE. Thus, they are best interpreted as lying on distinct mani- folds, and are best analyzed separately by LLE. 23. If neighbors correspond to nearby observations in time, then the reconstruction weights can be com- puted online (as the data itself is being collected) and the embedding can be found by diagonalizing a sparse banded matrix. 24. R. A. Horn, C. R. Johnson, Matrix Analysis (Cambridge

Univ. Press, Cambridge, 1990). 25. Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, H. van der Vorst, Eds., Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide (Society for Industrial and Applied Mathematics, Philadelphia, PA, 2000). 26. D. D. Lee, H. S. Seung, Nature 401, 788 (1999). 27. R. Tarjan, Data Structures and Network Algorithms, CBMS 44 (Society for Industrial and Applied Mathe- matics, Philadelphia, PA, 1983). 28. I. T. Jolliffe, Principal Component Analysis (Springer-

Verlag, New York, 1989). 29. N. Kambhatla, T. K. Leen, Neural Comput. 9, 1493

(1997). 30. We thank G. Hinton and M. Revow for sharing their unpublished work (at the University of Toronto) on segmentation and pose estimation that motivated us to “think globally, fit locally”; J. Tenenbaum (Stanford University) for many stimulating discussions about his work (4) and for sharing his code for the Isomap algorithm; D. D. Lee (Bell Labs) and B. Frey (University of Waterloo) for making available word and face data from previous work (26); and C. Brody, A. Buja, P. Dayan, Z. Ghahramani, G. Hinton, T. Jaakkola, D. Lee, F. Pereira, and M. Sahani for helpful comments. S.T.R. acknowledges the support of the Gatsby Charitable Foundation, the U.S. National Science Foundation, and the National Sciences and Engineering Research Council of Canada.

7 August 2000; accepted 17 November 2000