Distances, Dissimilarities, Similarities
And Clustering

By Daniel B. Carr

This is still an incomplete draft with some informal comments. Still, it can be useful if it provides broadened perspective or motivate a search for more information

1. Introduction

Many statistical analysis procedures are based on notions of distance between or dissimilarity of pairs of items. This resulting distance or dissimilarity matrices for all pairs have many uses. These include clustering, multidimensional scaling (MDS), other layout algorithms, minimal spanning tree construction, nonlinear dimension reduction and outlier detection. Our primary concern here is agglomerative clustering. The sequel to the class, STAT763, will address other applications. The brief sections below on other topics can be skipped if you are not curious.

The description below tends to use the general term dissimilarity matrix. Distances between three items, i, j and k must preserve the triangular inequality $d(i,k) \leq d(i,j) + d(j,k)$ while dissimilarities for three items do not. (See section 2 below for the properties of a distance metric.) In some situations the dissimilarity matrix discussed may be a distance matrix.

The relationship between pairs of items may be initially described in term of non-negative similarities. These can be transformed into dissimilarities. Two examples are

$$1 - \frac{s_{ij}}{\max(s_{ij})} \text{ and } \sqrt{(1-s_{ij})^2} \; for \; i \neq j.$$

1.1 Dissimilarity Matrices for Cases.

In the data table context we can apply dissimilarity producing functions to either pairs of cases or pairs of variables to get dissimilarity matrix. For the data table below cases correspond to rows and variable to columns.

	V1	V2	V3	V4	V5
C1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
C2	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}
C3	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}
C4	x_{41}	x_{42}	x_{43}	x_{44}	x_{45}
C5	x_{51}	x_{52}	x_{53}	x_{54}	x_{55}
C6	x_{61}	x_{62}	x_{63}	x_{64}	x_{65}
C7	x_{71}	x_{72}	x_{73}	x_{74}	x_{75}

Table 1. Original Data

If all the variables have the same units of measure a reasonable choice is to compute the distance between the cases using the variables as they are. If the variables have different units of

measure, then a common first step is to standardize each variable by subtracting its average and dividing by its standard deviation. Let z_{ij} be these unitless values.

	V1	V2	V3	V4	V5
C1	z_{11}	z_{12}	z_{13}	z_{14}	z_{15}
C2	z_{21}	z_{22}	z_{23}	z_{24}	z_{25}
C3	z_{31}	z_{32}	z_{33}	z_{34}	z_{35}
C4	z_{41}	z_{42}	z_{43}	z_{44}	z_{45}
C5	z_{51}	z_{52}	z_{53}	z_{54}	z_{55}
C6	z_{61}	z_{62}	z_{63}	z_{64}	z_{65}
C7	z_{71}	z_{72}	z_{73}	z_{74}	z_{75}

Table 2. Standardized columns with mean 0 and standard deviation 1

A common practice uses Euclidean distance to obtain a case distance matrix. Then the distance between cases i and j is

$$d(i,j) = \sqrt{\sum_{k=1}^{5}(z_{ik} - z_{jk})^2}.$$

Table 3 show a case distance matrix with zero on the diagonal

	C1	C2	C3	C4	C5	C6	C7
C1	0	d_{12}	d_{13}	d_{14}	d_{15}	d_{16}	d_{17}
C2	d_{21}	0	d_{23}	d_{24}	d_{25}	d_{26}	d_{27}
C3	d_{31}	d_{32}	0	d_{34}	d_{35}	d_{36}	d_{37}
C4	d_{41}	d_{42}	d_{43}	0	d_{45}	d_{46}	d_{47}
C5	d_{51}	d_{52}	d_{53}	d_{54}	0	d_{56}	d_{57}
C6	d_{61}	d_{62}	d_{63}	d_{64}	d_{65}	0	d_{67}
C7	d_{71}	d_{72}	d_{73}	d_{74}	d_{75}	d_{76}	0

Table 3: Case Distance Matrix

If we had not convert the values to be unitless, a simple change of scale for **one** variable, say from kilometers to millimeter could have a huge impact on the distance matrix.

When some variables are more important than others we can give them more weight using a weighted distance calculation. Let w_k , $k = 1{:}5$ be the weights for the 5 variables. Then

$$d(i,j) = \sqrt{\sum_{k=1}^{5} w_k(z_{ik} - z_{jk})^2}$$

In some situations we might consider centering or standardizing the **rows** of the data table before computing case distances. For this next example, focus attention on just three rows and three columns that are indicated in yellow below. (There may be instances when we may want to drop

variables from use in the clustering process and instances when we want tfocus on a subset of cases.

	V1	V2	V3	V4	V5
C1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
C2	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}
C3	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}
C4	x_{41}	x_{42}	x_{43}	x_{44}	x_{45}
C5	x_{51}	x_{52}	x_{53}	x_{54}	x_{55}
C6	x_{61}	x_{62}	x_{63}	x_{64}	x_{65}
C7	x_{71}	x_{72}	x_{73}	x_{74}	x_{75}

Case and Variable Focus for Data Table 1

Suppose we plot values for the three cases and time periods and obtain the results as shown in Figure 1 below.

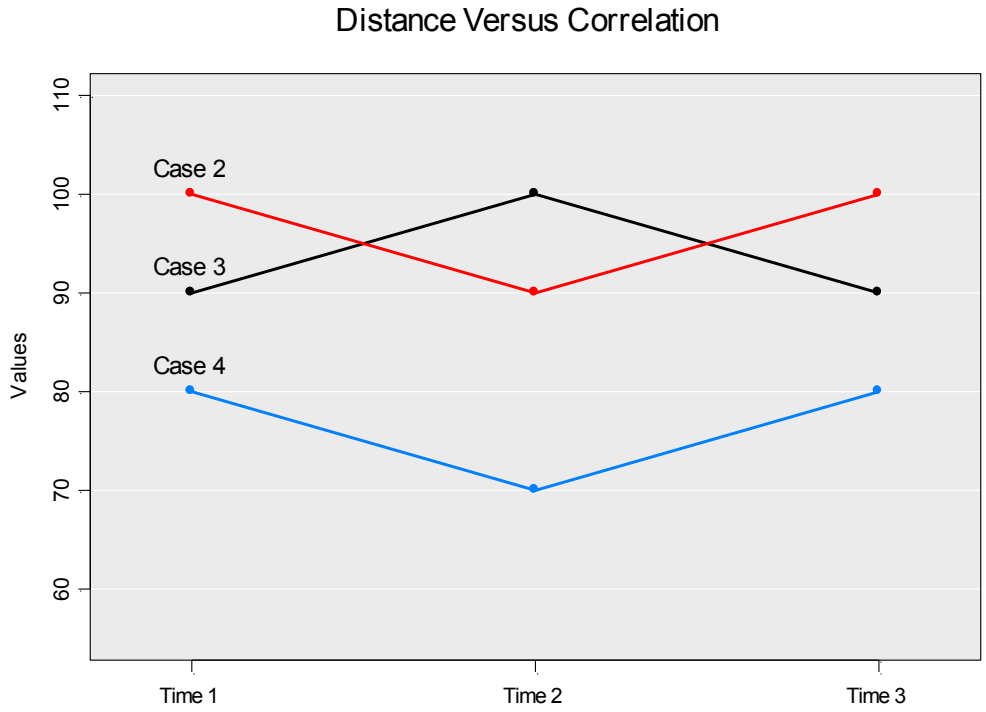

Figure 1: Data for 3 cases and first three variables

Table 4 gives the Euclidean distance matrix for cases 2, 3, and 4. We can see that cases 2 and 3 are close together while cases 2 and 4 are relative far apart. However, we can also see that cases 2 and 4 have the same pattern as also indicated by a correlation of 1 in Table 3 below, while cases 2 and 3 have a correlation of -1.

	C2	C3	C4
C2	0	$10\sqrt{3}$	$10\sqrt{12}$
C3	$10\sqrt{3}$	0	$10\sqrt{11}$
C4	$10\sqrt{12}$	$10\sqrt{11}$	0

Table 4: Euclidean Distance Matrix for Cases 2, 3 and 4

	C2	C3	C4
C2	1	-1	1
C3	-1	1	-1
C4	1	-1	1

Table 5: Correlation Matrix for Cases 2, 3 and 4

Consider clustering positively correlated cases or strong correlated cases. If we are correlation (or covariance) of case pairs as a foundation for clustering we need to make a further transformation to obtain non-negative dissimilarities to use in a dissimilarity matrix.

Table 6a gives the dissimilarity semi-metric calculated as 1- |correlation|. The values are all zero. This example is pathological because all the correlations are plus or minus 1. Note that for normal distribution a correlation of zero implies independence. We are used to thinking in terms of variables as being independent, but here we switching perspectives and entertain the idea of a case correlation of 0 as indicating independence under special conditions. Then for dissimilarity semi-metric of 1-|correlation| the values near 1 would be suggestive of independence.

	C2	C3	C4
C2	0	0	0
C3	0	0	0
C4	0	0	0

Table 6a: 1 - |correlation| Dissimilarity Matrix for Cases 2, 3 and 4

Table 6b below uses 1-correlation. A perfect negative correlation yields a dissimilarity of 1-(-1) = 2. A perfect positive correlation yields a dissimilarity of 1-1=0 that corresponds to the time series being the same up to a constant.

	C2	C3	C4
C2	0	2	0
C3	2	0	2
C4	0	2	0

Table 6b: Dissimilarity Matrix for Cases 2, 3 and 4

Which dissimilarity matrix should be used, Table 2, Table 6a, Table 6b or something quite different? Which is more important, the closeness of the points or the shape of the curves? The key point I want to make here is that we have choices in selecting a dissimilarity matrix.

In terms of the two correlations, I tend to use 1-|correlations| when I am ordering the variables of a correlation matrix. I want variables with the same information (whether expressed positively or negatively) to be close together. In the document analysis context, 1-correlation is often the choice.

In a document clustering context we could think of the documents as cases as documents and words as variables with cell entries c_{ij} being the word count for the i^{th} document and the j^{th} word. (Many descriptions use the transpose of this matrix with words indexing rows.)

Text mining example show some interesting variations. The R tm package supports text mining. The words are often stemmed words. Stemming remove prefixes and suffixes to reduce word variants to a common form. Punctuation such as a period or exclamation mark may be treated as word. Common practice removed very common words such as :the" and "a". A adjacent pairs of words (bigrams) or triple of words (trigrams) may define the columns. The list of n-grams can still be 10s of thousands long depending on the size and variety of documents. If there are 60,000 unique stemmed word in a document the number of bigrams could be very large.

For of corpus of document that data matrix of counts is often re-expressed as TFIDF statistics. The TF stands for the term frequency. The IDF stands for inverse document frequency. The TFIDF is the produce of the two component statistics. The term frequency statistics rescales the term count for a cell, dividing it by total count for the term across all document. The inverse document frequency statistic is log of the number of document divide by the number of documents with the term.

The normalized dot product of two document TFIFD vectors give the document similarity. Since the TFIDF matrix does not have negative values the dot product cannot be negative and 1 – normalized dot plot provides a dissimilarity measure. The dot product is the cosine of the angle between the pair of high dimensional vectors. If the dot product is 1, the angle is 0 degrees. If the dot product is 0, the angle is 90 degrees. With a dissimilarity matrix documents can be clustered. One of many application is to cluster conference abstracts and guide the authors to cluster corresponding bird-of-a-feather sessions.

The dissimilarly matrix, if not too large, can be fed into the singular value decomposition. The first few left eigenvectors provide new document coordinates so document can be shown as points in a scatterplot matrix and

1.2 Dissimilarity Matrices for Variables

As indicated above we can obtain dissimilarity measures for variable as well as cases. We can use 1-|correlations| or 1- correlation to compare the highlighted pair of rows in Table 1 with the

pair of highlighted pair columns in Tables 6a and 6b. As long as the units support meaningful calculations we can use normalized dot products (cosine of the angle between variables) as a measure of dissimilarity.

	V1	V2	V3	V4	V5
C1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
C2	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}
C3	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}
C3	x_{41}	x_{42}	x_{43}	x_{44}	x_{45}
C5	x_{51}	x_{52}	x_{53}	x_{54}	x_{55}
C6	x_{61}	x_{62}	x_{63}	x_{64}	x_{65}
C7	x_{71}	x_{72}	x_{73}	x_{74}	x_{75}

Table 6

2. Distance measures

By definition of distance, d(i,j), for two items involves the following requirements.

1) $d(i, i) = 0$
2) $d(i, j) \geq 0$ Non-negativity
3) $d(i , j) = d(j , i)$ Symmetry
4) $d(i , j) + d(j, k) \geq d(i , k)$ Triangular inequality

Note: Here $|x_{ik}|$ means absolute value and
$\|y\|$ means normalize the vector to have length 1.

Examples

Euclidean distance: $\sqrt{(x^T x)}$
Euclidean distance squared: $x^T x$

Mahalinobis distance: $x^T \sum^{-1} x$ where $\sum$ is covariance matrix (non-negative definite)

Chebyshev distance $\max_k \left| x_{ik} - x_{jk} \right|$

City block (Manhattan) $\text{Sum}_k \left| x_{ik} - x_{jk} \right|$
Great arc distance for latitude and longitude

Fixes for extreme values

Winsorize: replace all really extreme values by lesser extreme value.
Replace all variable values by normal scores, etc.

Distance between distributions and multivariate distributions

In statistics there are several ways to assess the distance between continuous univariate distributions. In earth sciences and many complex application areas we are often interested in the distance between multivariate distributions.

I learned about one very useful distance measure from approach from Amy Braverman, JPL when she gave a talk at GMU. It is called the earth move distance in image processing and the Wasserstein metric and an other names other literature. See Wikipedia. The application was to assess the expected distance between earth grids. Each grid cell had a set of summary multivariate vectors with weights. Each vector has values for temperature, water vapor, and cloud fraction at multiple altitudes. Each earth grid cell has a different number of summary vectors. Once was has the distance matrix for earth grid cells we could use hierarchical clustering, color the clusters and on map of the earth and see patterns.

3. Dissimilarity semi-metrics

The dot product of normalized vectors give the cosine of the angle between vectors.

Cosine (0) = 1. Thus for identical vectors, the angle is zero and dissimilarity calculation below, 1-cos(0), yields 0 as is desired.

1 - Cosine: $1 - x \cdot y / (\|x\| \|y\|)$ [0, 2]
or $1 - \left| x \cdot y / (\|x\| \|y\|) \right|$ [0, 1]
or $\left| \arccos (x \cdot y / (\|x\| \|y\|)) \right|$ [0, 180] or [0, pi]
Correlation : $1 - \text{cor}(x, y)$ [0, 2]
$1 - \left| \text{cor}(x, y) \right|$ [0, 1]

Note that for the normal (Gaussian) distribution, zero correlation implies independence. This does not hold in general. Still my choice in term of correlation is $1-|\text{cor}(x, y)|$ since in MDS it tends to put nearly independent variables far apart.

4. Transformations to handle different kinds of variables

In general sections 1) and 2) were intended to handle interval scaled variables (continuous measurement on a roughly linear scale). There are ways to transform other kind of variables so the above methods become applicable

4.1 Continuous and Discrete Ordinal Variables

Replace x_{ik} by their rank r_{ik} when sorted.

Transform into [0 1] using $(r_{ik}-1)/(\max(r_{ik})-1)$

4.2 Vector of Nominal Variables

1) Number of variables taking on different values for I and J ／ Number of variables

2) Mutual Information (see Section 6)

4.3 Vector of symmetric binary variables: handle like 4.2

4.4 Vector of asymmetric binary variables

Here one value is more important than the other. The important variable is coded 1 and unimportant variable is coded 0.

Number mismatches ／ (# variables - # variables both 0)

In one application the objective was to assess the distance between bird species based on where they live. The nation was divided into 13000 regions. For each bird species the value 1 for a region meant the species was observed or inferred to live in the region. The value 0 meant the species was absent. The distance between species would often look close due to the large number of places where neither species lived. To care this to extreme, the distance between species would be even closer if little regions across all of the Atlantic and Pacific oceans were included for U.S. land-based bird species. The calculation that was used was restricted to regions in which at least one of the two species was present**.**

4.5 Ratio scaled variables

These have positive continuous values on a nonlinear scale such as an exponential scale.

i) take logs

ii) treat as continuous ordinal

iii) use as is (not recommended)

4.6 Mutual Information for blocks of counts

As an example consider the distance between documents based on the stemmed words that they contain. As a preparatory step transform the vector of word counts for each document into a mutual information vector.

The mutual information is calculated as follows:

Let N be the total count of (stemmed) words in all documents.
Let Cik be the count for document i and word k.

Let Pik = Cik/N be the two-way cell probability estimates

Let Pi· = Ci·/N and P·k = C·k/N be the row and column margin probability estimates.

The mutual information for document i and word k is Mik = $\log_2$ (Pik /(Pi· * P·k))

The Preibe et al. 2004 suggest a discounting factor for infrequent words:

DFik = Cik/(Cik + 1) * (N*min(P·k,Pi·)/ (1+N*min(P·k, Pi·)))

The discounted mutual information vector is

M*ij = Mij*DFij

Priebe et al. (2004) use the cosine semi-metric to obtain dissimilarities between document i and j.

5. More complex scenarios

5.1 Mixtures across multiple groups of variables

Use a weighted linear combination of distances for groups of variables with the weights summing to 1. For example one can mix geographic distance between a pair of cases with the distance between time series for the pair of cases. A challenge is to decide what to weight most heavily.

5.2 Combining different kinds of cases (for example Terrorist and Events)

When the distance (or dissimilarity) matrix is created from two or more sets of different kinds items, the matrix has a block structure. For example there might be terrorists and events. The terrorist distances (or dissimilarities) would comprise one block. The events distances (or

dissimilarities) would comprise another. The distances (or dissimilarities) across the two sets create two additional blocks that are transposes of each other.

To experiment with different views, one can multiple the different blocks by different weight factors. One weight can be taken to be 1. In one view one might want to see which terrorists are close to other terrorists. In another view, interest might be about which terrorists are close to which events. Yet another third view can emphasize which events are similar.

I have long wanted to layout the terrorist and events in two different planes with the front one translucent. I wants the graphics in the movie, Minority Report, extended to two planes. Lines between the terrorist and event planes would show indicate strength of the connection. Michael Trosset, an expert in MDS layouts said he knew to do the layout. It would only take a little time write it down. I pestered a few times to do so, but to no avail.

5.3 Directional semi-metrics

(Will add this later)

6. Finding good subspaces of variables for clustering

GMU experts include Dr. Carlotta Domeniconi and Dr. Daniel Barbara both in the Computer Science department. I am know relatively little. My old class example is based the paper by

Guo, D., Gahegan, M., Peuquet, D., & MacEachren, A. 2003, Breaking Down Dimensionality: An Effective Feature Selection Method for High-Dimensional Clustering. *Workshop on Clustering High Dimensional Data and its Applications, the Third SIAM International Conference on Data Mining*, San Francisco, CA, May 1-3.

The example steps illustrated in an assignment are:

Step 1. Define a rectangular grid to convert a scatterplot into a two-way table of counts.

They recursively split the range of each variable based on the mean. They would like like to have 35 counts in each two-way cell. Let s be the number or 1-dimensional splits. They suggested s = floor(.5*log2 (totalCounts/35))
Their splitting choice is apparently based on speed.

As a slower alternative in R we could readily find
nbin = floor (sqrt(totalCounts/35)) roughly equal count

Step 2. Find the table margin probability $Pi\cdot$ and entropy $Ei\cdot$ for each row.
$Ei\cdot$ = - Sum(p_j*log(p_j,base=2)). The probs p_j are from just the row itself.
Sum the dot product of $Pi\cdot$ and $Ei\cdot$ getting CE(X|Y)
Repeat for the columns getting CE(X|Y)
Let CEM (X,Y) = max(CE(X|Y),CE(Y|X))

Step 3. Do this for all the pairs of variables

Step 4. Sort the variables based use SVD (singular value decomposition) or minimal spanning tree traversal.
(See early discussion of minimal spanning tree traversal for variables)
Get the traversal order from mstree()

Step 5. Plot a color CEM matrix using the sorted variables order and look for blocks of Low CEM values. These give the subspace of interest.
(Graphically Guo et al put CEM below the diagonal an correlation about the diagonal in the color matrix and use to two different color scales.

There is an illustrative R script in an assignment.

7. Short comments on distance matrix uses

7.1 Case and variable reduction

We can cluster cases or variables. This provides a basis for case reduction and dimension reduction. That is case clusters and variable clusters can be replaced by a smaller number of representative cases or variables or by other forms of statistical summaries. In the random forest assignment this and example of produce prototypes that can represent several of it nearest neighbors.

7.2. Dimension reduction via multidimensional scaling.

Another use is to produce coordinates for views of higher dimensional data in lower dimensions in ways that strive to preserve the interpoint distances between all the pairs of points. R has cmdscale() function for the classical multidimensional scaling approach. Basically the centers the distance matrix appropriately, feed this to the singular value decomposition function and use the first 2, 3 or more eigenvectors as the coordinates to plots. (Two function in the MASS package, isoMDS() and sammon() provide non-metric multidimentional scaling.)

Sometimes a low dimensional views will reveal low dimensional structure is embedded in high dimensional space. Sometimes MDS scaling or other layouts can produce strange clusters in low dimensions. That is, the items in appearing to be clusters in low dimensions can be coming from quite different parts of high dimensional space but forced together because other points are even further apart.

Perhaps we should not let the MDS or other algorithms do all the compromising for us. Our focus is often on items that are very close (or similar) or of intermediate proximity and we may not care that much about accurate portrayal of large distances as long they are large enough to escape our attention. This leads to thoughts about reducing the large distances in the interpoint distance matrix.

7.3 Spring models and other layouts

Given distances or dissimilarities, many layout approaches are available. Spring models are fairly popular. There are space filling layouts for hierarchical clustering (see Eick and Wills). Ru Sun, one of my Ph.D. students and I have developed some algorithms to place cases in round regions in 2-D and 3-D rather than in traditional rectangular regions.

Lisong Sun described and interesting containment tree layout with illurstrates this with some very appealing graphics. See L. Sun, S Smith and T.P. Caudel [2003]. *A Low Complexity Recursive Force-Directed Tree Layout Algorithm Based on the Lennard-Jones Potential*. UNM Technical Report: EECE-TR-03-001. This is likely published elsewhere possible with additional material.

A fairly obvious pragmatic approach layout a representative vector (or a few) from each of the major clusters. Then items in the cluster can be laid out relative to the representative vectors. Layouts are a topic for an extended paper and a fun area for experimentation.

Outliers can be readily evident in d-matrix (d- stands for distance or dissimilarity) when the matrix is not too large to visualize (Lukens 2004). My matrices are too large. This topic will be fleshed out later.)

7.4 Seriation methods and variable ordering

In many cases is desirable to establish a rank order for variables. (The topic of rank order often appears under the label "seriation.") For example how should one order the variables for a parallel coordinate plot. My common approach uses 1-|correlation| to obtain a dissimilarity matrix for the variables. Then I used to use

1) MDS algorithm to get pseudo coordinates
2) feed the pseudo coordinates the Splus minimal spanning tree program, and
3) use the breadth tree traversal order to determine the variables order.

7.5 Point to point Distance, nonlinear dimension reduction methods and Manifold Learning

What is the most relevant path between points and how long is it? When the Euclidean distance path between two points goes through regions of space with no data, the path may a poor basis for assessing distance.

A better path may well be the shorted path that goes from neighbor to neighbor to get between two points. We will return in the context of using neighbor to neighbor geodesic distance (ISOMAP) and diffusion paths through the nonlinear manifolds.

Since 2000, several papers have addressed the ideas of geodesic paths (shortest distance in curved space) and nonlinear dimension reduction.

References

Belkin, M. and P. Niyogi. 2003. "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," Neural Computation, June 2003; 15 (6):1373-1396.

Donoho, D. L. and C. Grimes. 2003. "Hessian Eigenmaps: Locally Linear Embedding Techniques for High-dimensional Data, Proceeding of the National Academy of Sciences U S A. 2003 May 13; 100(10): 5591:5596.

S. Lafon. 2004. Diffusion maps and geometric harmonics,Ph.D. dissertation, Yale University May 2004.

Roweis, S. and L. Saul. 2000. "Nonlinear dimensionality reduction by locally linear embedding," Science, 290(5500), 2323:2326.

Tenenbaum, J. B. , V. de Silva, and J. C. Langford. 2000."A global geometric framework for nonlinear dimensionality reduction," Science 290(5500), 2319:2323.

Izenman 2008 has a chapter in his book devoted to this topic.