Distances, Dissimilarities,
And Selected Applications

By Daniel B. Carr
Draft: November 10, 2004

1. Introduction

Many statistical analysis procedures are based on notions of distance or dissimilarity for pairs of items. These include clustering, multidimensional scaling (MDS), other layout algorithms and outlier detection methods. In the data table context we can apply methods to cases or variables.

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.

We can plot cases or variables in 2D or 3-D scatterplots or other relatively low dimensional layouts that are intended to be cognitively accessible. Sometimes a low dimensional structure is embedded in high dimensional space and a low dimensions view will provide reasonably faithful rendering of the interpoint distances in high dimensions. Sometimes MDS scaling or other layouts can produce strange clusters in low dimensions. That is, the items in the displayed clusters are 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.

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.

Given distances or dissimilaries, many layout approaches are available. Spring models are fairly popular. There are space filling layouts for hierarchical clustering (see Eick and Wills). One of my Ph.D. students and I have developed some algorithms. 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.)

Data Table 1 with 7 cases and 5 variables provide a starting place to introduce some of the methods and issues related to distances and dissimilarities. First consider the distance or dissimilarity between pairs of cases based when focusing our attention on variables 1, 2, 3. The light blue highlighting indicates the data used in the assessment for cases 2, 3 and 4.

	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}

Data Table 1

Suppose variables 1, 2, and 3 represent times and the values for cases 2, 3, and 4 are as shown in Figure 1 below

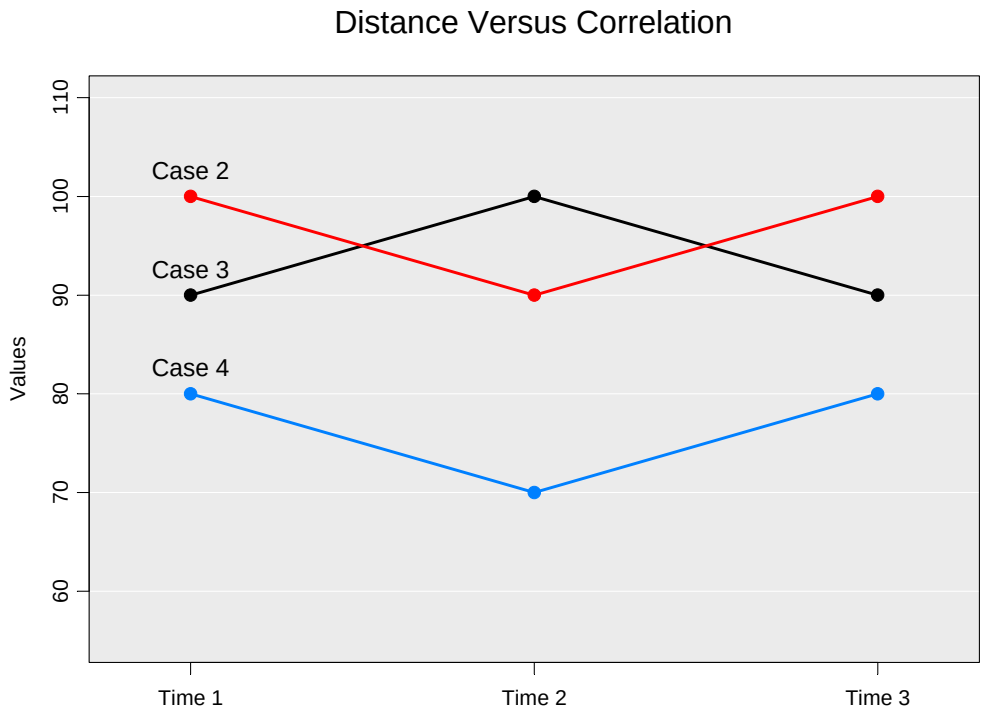

Figure 1: Data for 3 cases.

Table 2 gives the distance matrix for cases 2, 3, and 4. Table 3 gives the correlation matrix for the same cases. We can see at a glance that cases 2 and 3 are similar in terms of Euclidean distance but negatively correlated. Cases 2 and 4 are pretty far apart in terms of Euclidean distance but perfectly correlated. Cases 3 and 4 are pretty far apart in terms of Euclidean distance and negatively correlated.

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

Table 4a gives the dissimilarity semi-metric calculated as 1- |correlation|. The value area all zero. This example is pathological because all the correlations are plus or minus 1. Note that with the normal distribution a correlation of zero implies independence. As dissimilarity of 1 would then indicated the variables are not relates while a dissimilarity of 0 indicates they are perfectly related.

Table 4b uses 1-correlation. A perfect negative correlation yields a dissimilarity of 2. A perfect positive correlation yields a dissimilarity of 0. Thus the time series is the same up to a constant.

Which should be used Table 2, Table4a, Table 4b or something quite different? Which is more important, the closeness of the points or the shape of the curves? 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 terms of providing a linear
Order of variab
, I prefer 1-|correlWith the normal distribution a correlation of zero implies independence.

May be the dissimilarity metric should be 1 – correlation. What data consideration might be used to adapt the distances or dissimilarities? For example,
Suppose the times in Figure 1 were not equally spaced, and the variables assess close in time had almost identical values? Perhaps the two similar variables should be averages or individually down weighted in some way become the information seems redundant.

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

Table 4a: Dissimilarity Matrix for Cases 2, 3 and 4

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

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

As indicated above we can obtain distance and dissimilarity measures for variable as well as cases. Compare the highlighted pair of rows in Table 1 with the pair of highlighted pair columns in Table 5. 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. It is then possible to cluster variables, layout variables, and identify outlier variables.

	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 5

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 use

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

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 $\max_k |x_{ik} - x_{jk}|$

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

Fixes for extreme values

Winsorize

Replace all variable values by normal scores

Etc.

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 $1 - |x \cdot y/(\|x\| \|y\|)|$ [0, 1]
or $|\arccos(x \cdot y/(\|x\| \|y\|))|$ [0, 180] or [0, pi]

Correlation : $1 - \text{cor}(x, y)$ [0, 2]
$1 - |\text{cor}(x, y)|$ [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 $P_{i\cdot} = C_{i\cdot}/N$ and $P_{\cdot k} = C_{\cdot k}/N$ be the row and column margin probability estimates.

The mutual information for document i and word k is $M_{ik} = \log_2 (P_{ik} /(P_{i\cdot} * P_{\cdot k}))$

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

$DF_{ik} = C_{ik}/(C_{ik} + 1) * (N*\min(P_{\cdot k},P_{i\cdot})/ (1+N*\min(P_{\cdot k}, P_{i\cdot})))$

The discounted mutual information vector is

$M^*_{ij} = M_{ij}*DF_{ij}$

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.

5.3 Directional semi-metrics

(Will add this later)

6. Finding good subspaces of variables for clustering

(Various approaches have their limitations. Limitations of this approach will be discussed later.)

A data table sometimes contains irrelevant data for the task at hand, or noise variables.
Guo, Gahegan, Peuquet, and MacEachren (2003?) provide an approach to locating subspaces (identifying sets of variables) of continuous variables in which there is a fair to high amount of clustering.

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 Splus we could readily find
nbin <- floor (sqrt(totalCounts/35)) roughly equal count

Step 2. Find the table margin probability Pi· and entropy Ei· for each row.
Ei· = - Sum(p_j*log(p_j,base=2)). The probs p_j are from just the row itself.
Sum the dot product of Pi· and Ei· 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 Splus script in an assignment.