Interpreting Similarity

Michael W. Trosset
Department of Statistics
Indiana University

Supported by a National Security Science & Engineering Faculty Fellowship awarded to Carey E. Priebe (Johns Hopkins University).
Proximities are real numbers that convey pairwise information about the extent to which two objects resemble each other. If larger proximity indicates that objects are less alike, then we say that the proximities measure dissimilarity and denote the proximity of pair $i \sim j$ by $\delta_{ij} = \delta_{ji}$. If larger proximity indicates that objects are more alike, then we say that the proximities measure similarity and denote the proximity of pair $i \sim j$ by $\gamma_{ij} = \gamma_{ji}$. By convention, one usually requires proximities to be nonnegative. Some authors further require $\gamma_{ii} = 1$.

Proximities may be measured directly; more often, they are computed from feature vectors, say $y_1, \ldots, y_n \in \mathbb{R}^q$. 
Motivating Example

Suppose that we survey \( n = 4 \) film buffs, observing whether or not each individual has or has not seen each of the \( q = 6 \) feature films directed by Sergio Leone:

- *A Fistful of Dollars* 1964
- *For a Few Dollars More* 1965
- *The Good, the Bad, and the Ugly* 1966
- *Once Upon a Time in the West* 1968
- *Duck, You Sucker!* 1972
- *Once Upon a Time in America* 1984

Let \( y_{ik} = 1 \) if individual \( i \) has seen film \( k \) and let \( y_{ik} = 0 \) if s/he has not. Then we might observe the following data matrix:

\[
Y = \begin{bmatrix}
1 & 1 & 1 & 1 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 1 & 0 & 0 & 1
\end{bmatrix}.
\]
We might measure the similarity of pairs of film buffs by computing simple matching coefficients. The matching coefficient for pair \( i \sim j \) is the proportion of films that both \( i \) and \( j \) either have or have not seen. Thus,

\[
\gamma_{ij} = \frac{1}{6} \sum_{k=1}^{6} \left\{ \begin{array}{ll} 
1 & \text{if } y_{ik} = y_{jk} \\
0 & \text{if } y_{ik} \neq y_{jk} 
\end{array} \right\}
\]

and

\[
\Gamma = [\gamma_{ij}] = \frac{1}{6} \begin{bmatrix}
6 & 5 & 4 & 2 \\
5 & 6 & 5 & 3 \\
4 & 5 & 6 & 2 \\
2 & 3 & 2 & 6 
\end{bmatrix}.
\]
With equal plausibility, we might measure the dissimilarity of individuals $i$ and $j$ by computing the proportion of films that one of the individuals has seen and the other has not, i.e., $\delta_{ij} = 1 - \gamma_{ij}$. Doing so results in the dissimilarity matrix

$$
\Delta_2 = [1 - \gamma_{ij}] = \frac{1}{6} \begin{bmatrix}
0 & 1 & 2 & 4 \\
1 & 0 & 1 & 3 \\
2 & 1 & 0 & 4 \\
4 & 3 & 4 & 0
\end{bmatrix}.
$$

It would seem that we should be completely indifferent to describing the proximities between our film buffs by $\Gamma$ or by $\Delta_2$. For example, there is no evident difference between saying that individuals 1 and 3 match on 4 of 6 films ($\gamma_{13} = 4/6$) or that they mismatch on 2 of 6 films ($\delta_{13} = 2/6$).
Alternatively, notice that the possible feature vectors for this example lie in $\{0, 1\}^6 \subset \mathbb{R}^6$. Hence, we might also measure the dissimilarity of individuals $i$ and $j$ by computing the Euclidean distance between their feature vectors. Because $y_{ik}, y_{jk} \in \{0, 1\}$,

$$d_{ij}^2 = \|y_i - y_j\|^2 = \sum_{k=1}^{6} (y_{ik} - y_{jk})^2 = \sum_{k=1}^{6} \left\{ \begin{array}{ll} 0 & \text{if } y_{ik} = y_{jk} \\ 1 & \text{if } y_{ik} \neq y_{jk} \end{array} \right\} = 6 \left( 1 - \gamma_{ij} \right).$$

Hence,

$$\Delta_1 = \left[ (1 - \gamma_{ij})^{1/2} \right]$$

is also a plausible dissimilarity matrix.

The dissimilarity matrix $\Delta_1$ is the matrix of Euclidean distances between the (scaled) feature vectors. It turns out that the dissimilarity matrix $\Delta_2$ cannot be realized as the matrix of pairwise Euclidean distances for any configuration of feature vectors.

Should we regard $\delta_{ij} = 1 - \gamma_{ij}$ as a measure of dissimilarity or as a measure of squared dissimilarity?
The notion of dissimilarity is motivated by the formal concept of distance. If $\Delta$ is a matrix of pairwise distances, then the following properties necessarily obtain.

1. For each $ij$, $\delta_{ij} \geq 0$, i.e., $\Delta$ is nonnegative.

2. For each $i$, $\delta_{ii} = 0$, i.e., the diagonal entries of $\Delta$ vanish, i.e., $\Delta$ is hollow.

3. For each $ij$, $\delta_{ij} = \delta_{ji}$, i.e., $\Delta = \Delta^t$, i.e., $\Delta$ is symmetric.

We shall refer to these properties as the *delta properties* and use them to provide a formal definition of dissimilarity.

The $n \times n$ matrix $\Delta$ is a *dissimilarity matrix* iff it satisfies the delta properties, i.e., iff $\Delta$ is hollow, nonnegative, and symmetric.
The mathematical concept of distance provides a natural model of dissimilarity, but how might we model similarity? Here is a formal definition:

The \( n \times n \) matrix \( \Gamma = [\gamma_{ij}] \) is a similarity matrix iff \( \Gamma \) is symmetric and \( \gamma_{ii} \geq \gamma_{ij} \geq 0 \) for each \( ij \).

We write

\[
\|y_i - y_j\|^2 = \langle y_i - y_j, y_i - y_j \rangle = \|y_i\|^2 - 2 \langle y_i, y_j \rangle + \|y_j\|^2
\]

and impose two restrictions on the feature vectors:

1. Assume that the \( y_i \) lie in the positive orthant of \( \mathbb{R}^q \), so that \( \langle y_i, y_j \rangle \geq 0 \).

2. Assume that the \( y_i \) lie on the unit sphere, so that \( \|y_i\|^2 = \|y_j\|^2 = 1 \).

Then \( \|y_i - y_j\|^2 = 2 - 2\langle y_i, y_j \rangle \), and \( \langle y_i, y_j \rangle \) varies inversely with \( \|y_i - y_j\|^2 \).
The preceding calculations motivate the use of the Euclidean inner product as a measure of similarity. Notice the following consequences of restricting the feature vectors to the unit sphere:

- \( \gamma_{ii} = \langle y_i, y_i \rangle = \|y_i\|^2 = 1 \), a common assumption.

- Because \( \|y_i\| = \|y_j\| = 1 \),

\[
\langle y_i, y_j \rangle = \frac{\langle y_i, y_j \rangle}{\|y_i\| \cdot \|y_j\|} = \cos \alpha_{ij},
\]

where \( \alpha_{ij} \) is the angle between \( y_i \) and \( y_j \). Hence, \( \gamma_{ij} = \langle y_i, y_j \rangle \) is widely known as \textit{cosine similarity}.
Without restricting the feature vectors to the unit sphere, it is hard to argue that the Euclidean inner product measures similarity. For example, let

\[ y_i(t) = \begin{bmatrix} 1 + t \\ 1 \end{bmatrix} \quad \text{and} \quad y_j(t) = \begin{bmatrix} 1 \\ 1 + t \end{bmatrix} \]

for \( t \geq 0 \). Then

\[ \|y_i(t) - y_j(t)\|^2 = 2t^2 \quad \text{and} \quad \langle y_i(t), y_j(t) \rangle = 2 + 2t, \]

both of which increase with \( t \).
Euclidean Representations of Proximity

Euclidean distance is a type of dissimilarity that most people find fairly intuitive, and a great many statistical techniques have been developed with Euclidean distance in mind. For these reasons, one often endeavors to approximate non-Euclidean dissimilarities with Euclidean distances. The construction of such approximations is variously described as embedding, ordination, or multidimensional scaling.

Notice that it is easy to embed inner products. Any $B \geq 0$ can be embedded by first computing $B = XX^t$, then setting $x_i^t$ equal to row $i$ of $X$. Furthermore, it is easy to obtain pairwise inner products from pairwise Euclidean distances. These observations motivate classical multidimensional scaling (CMDS).

Theorem: Let $\Delta = [\delta_{ij}]$ be a dissimilarity matrix and $\Delta_2 = [\delta^2_{ij}]$. Let $P = I - ee^t/n$. Then $\Delta$ is a matrix of pairwise Euclidean distances (EDM-1) iff

$$B = \tau (\Delta_2) = -\frac{1}{2} P \Delta_2 P \geq 0.$$ 

Furthermore, if $B \geq 0$, then $B = XX^t$ entails $D(X) = \Delta$. 
Classical Multidimensional Scaling

1. Compute $\Delta_2 = [\delta_{ij}^2]$ and $B = \tau(\Delta_2)$.

2. Compute $\lambda_1 \geq \cdots \geq \lambda_d$, the $d$ largest eigenvalues of $B$, and corresponding orthonormal eigenvectors $u_1, \ldots, u_d$.

3. For $i = 1, \ldots, d$, let $\bar{\lambda}_i = \max(\lambda_i, 0)$.

4. Let $\bar{\lambda}_i = \sigma_i^2$ and $X = \begin{bmatrix} \sigma_1 u_1 & \cdots & \sigma_d u_d \end{bmatrix} = \begin{bmatrix} x_1^t \\ \vdots \\ x_n^t \end{bmatrix}$.

Then $X$ is centered, its Cartesian coordinate axes are its principal component axes, and $XX^t$ is the rank-$d$ inner product matrix that best approximates $B$.

If the $\delta_{ij}$ are the pairwise Euclidean distances between feature vectors in $\mathbb{R}^q$, then CMDS is principal component analysis. Because the principal component representation is extracted from the matrix of pairwise inner products, CMDS is a kernel method.
CMDS embeds fallible distances ($\Delta$) by transforming them to fallible inner products ($B$), then approximating $B$ with Euclidean inner products. In principle, one should interpret the resulting configuration in terms of its pairwise inner products. No one does. Distances are easier to interpret than inner products—especially when the inner products have been centered. Even experts frequently describe CMDS as a technique that approximates dissimilarity with Euclidean distance.

A more intuitively appealing approach is to optimize the approximation of dissimilarity by Euclidean distance, e.g., by minimizing

$$\sigma(X) = \sum_{i<j} w_{ij} [d_{ij}(X) - \delta_{ij}]^2,$$

where $d_{ij}(X) = \|x_i - x_j\|$ and $w_{ij} \geq 0$. 
Embedding Similarity

If $\Gamma \geq 0$, then one can embed $\Gamma$ by computing $\Gamma = XX^t$. If $\gamma_{ii} = 1$, then $\gamma$ is cosine similarity in the Euclidean representation of $\Gamma$. Many popular measures of similarity can thus be represented as cosine similarity in $\mathbb{R}^d$, e.g., the heat kernel,

$$\Gamma = \left[ \exp \left( -c \| y_i - y_j \|^2 \right) \right].$$

In practice, however, the advantages of a principal component representation are so compelling that one rarely embeds by computing $\Gamma = XX^t$. Rather, the standard practice is to embed by computing $P \Gamma P = XX^t$.

Why?
If $B = [b_{ij}] \geq 0$ is a matrix of pairwise inner products, then

$$D_2 = [d^2_{ij}] = \text{diag}(B) e^t - 2B + e \text{diag}(B)^t = \kappa(B)$$

is the corresponding matrix of squared Euclidean distances.

If $b_{ii} = 1$, then $\kappa$ simplifies to $d^2_{ij} = 2[1 - b_{ij}]$.

To construct a principal component representation from a fallible $\Gamma$, first apply the “standard transformation” $\Delta_2 = \kappa(\Gamma)$ to obtain fallible squared distances, then proceed as in CMDS. Mathematically,

$$B = \tau(\kappa(\Gamma)) = P\Gamma P,$$

thereby simplifying the necessary calculations. However, while it is standard practice to obtain $B$ from $\Gamma$ by double centering $\Gamma$, $\kappa$ may not be the most appropriate transformation from similarity to (squared) dissimilarity. Furthermore, it easier to visualize and interpret distances and dissimilarities than to visualize and interpret centered inner products and similarities. Accordingly, we favor first transforming $\Gamma$ to a suitable $\Delta$, then embedding $\Delta$. 
Cosine Similarity

Imagine a corpus of $n$ documents about dogs and/or wolves. Let $y_{i1}$ denote the number of times the word *dog* or *dogs* appears in document $i$ and let $y_{i2}$ denote the number of times the word *wolf* or *wolves* appears in document $i$. These word counts impart some information about document content. For example, documents about how dogs evolved from wolves may have comparable values of $y_{i1}$ and $y_{i2}$, whereas documents about training dogs to heel will tend to have $y_{i1} \gg y_{i2}$ and documents about reintroducing wolves to Yellowstone National Park will tend to have $y_{i2} \gg y_{i1}$.

To remove the effect of document length, we construct equivalence classes defined by $y_{i1}/y_{i2}$. Geometrically, the equivalence classes are rays in the positive orthant of $\mathbb{R}^2$ that emanate from the origin. A natural measure of dissimilarity between two such equivalence classes is the angle between their corresponding rays.

Equivalently, let $z_i = y_i/\|y_i\|$, let $\alpha_{ij}$ denote the angle between $y_i$ and $y_j$, and set $\delta_{ij} = \alpha_{ij}$. The $z_i$ lie on the unit sphere, and $\alpha_{ij}$ is also the geodesic ("great circle") distance between $z_i$ and $z_j$. 
The angle between $z_i$ and $z_j$ is

$$\alpha_{ij} = \arccos \langle z_i, z_j \rangle = \arccos \frac{\langle y_i, y_j \rangle}{\|y_i\| \cdot \|y_j\|}.$$ 

Because $\cos$ is decreasing on $[0, \pi/2]$,

$$\gamma_{ij} = \cos \alpha_{ij} = \frac{\langle y_i, y_j \rangle}{\|y_i\| \cdot \|y_j\|}$$

is a measure of similarity. Given the preceding rationale for computing $\gamma_{ij}$, the correct transformation from cosine similarity to dissimilarity is $\delta_{ij} = \arccos \gamma_{ij}$. Oddly, this transformation is rarely used.

In practice, $1 - \gamma_{ij}$ is often used to measure dissimilarity. Notice, however, that $\gamma_{ii} = 1$ and therefore $2(1 - \gamma_{ij})$ is the standard transformation from similarity to squared dissimilarity.
Furthermore, the trigonometric power series expansion

\[ 1 - \cos \alpha = \frac{\alpha^2}{2!} - \frac{\alpha^4}{4!} + \frac{\alpha^6}{6!} - \cdots, \]

reveals that \( 2(1 - \cos \alpha) \sim \alpha^2 \) for small \( \alpha \). This revelation suggests that \( 1 - \cos \alpha_{ij} \) is not expressed in the appropriate units. The obvious remedy is to take square roots, resulting in dissimilarities

\[ \hat{\delta}_{ij} = \sqrt{2 \left(1 - \cos \alpha_{ij}\right)}. \]

In fact, \( \hat{\delta} \) has a nice geometric interpretation. Some elementary trigonometric calculation shows that

\[ \|z_i - z_j\|^2 = (1 - \cos \alpha_{ij})^2 + \sin^2 \alpha_{ij} = 1 - 2 \cos \alpha_{ij} + \cos^2 \alpha_{ij} + \sin^2 \alpha_{ij} = 2 - 2 \cos \alpha_{ij}, \]

so \( \hat{\delta}_{ij} = \|z_i - z_j\| \) is the \textit{chordal distance} between \( z_i \) and \( z_j \), and the standard transformation \( \hat{\Delta}_2 = \kappa(\Gamma) \) computes squared chordal distances.
Embedding Cosine Similarity

Consider using CMDS to embed $\Delta = [\alpha_{ij}]$ and $\hat{\Delta} = [\hat{\delta}_{ij}]$.

Embedding $\Delta$ constructs a $d$-dimensional Euclidean representation of $n$ points on the unit sphere in $\mathbb{R}^q$.

In contrast, $\hat{\delta}_{ij}$ is the $q$-dimensional Euclidean distance between $z_i$ and $z_j$, and therefore $\hat{\Delta}$ is EDM-1. Hence, embedding $\hat{\Delta}$ by CMDS is equivalent to performing PCA on $z_1, \ldots, z_n \in \mathbb{R}^q$. As the following example illustrates, projecting $z_1, \ldots, z_n$ into a $d$-dimensional hyperplane may distort the data in ways that are inconsistent with the rationale that is usually invoked to justify the use of cosine similarity.
Example

Let $y_i \in \mathbb{R}^2$ be such that

$$z_i = \frac{y_i}{\|y_i\|} = \left( \cos \frac{i \pi}{20}, \sin \frac{i \pi}{20} \right)$$

for $i = 1, \ldots, 19$. The $z_i$ are equally spaced on the unit circle with pairwise angles

$$\alpha_{ij} = \frac{|i - j| \pi}{20}. $$
The matrix of pairwise angles, $\Delta = [\alpha_{ij}]$, is also the matrix of pairwise Euclidean distances of $n = 19$ equally spaced points on the real line; hence, $\Delta$ is EDM-1 with embedding dimension $p = 1$ and the 1-dimensional principal component representation of $\Delta$ comprises the equally spaced points

$$x_i = \frac{i - 10 \pi}{20}.$$ 

In contrast, $\hat{\Delta} = [\|z_i - z_j\|]$ is EDM-1 with embedding dimension $p = 2$. The first principal component axis passes through the centroid of the $y_i$; by symmetry, its slope is $-1$. Projecting $z_i$ into this 1-dimensional hyperplane, we obtain the 1-dimensional principal component representation of $\hat{\Delta}$,

$$\hat{z}_i = \cos \left( \frac{i \pi}{20} + \frac{\pi}{4} \right) = \cos \left( \frac{i + 10 \pi}{20} \right).$$
Simple Matching

Given \( y_1, \ldots, y_n \in \{0, 1\}^q \), let \( \Gamma = \gamma_{ij} \) denote the similarity matrix of simple matching coefficients, i.e., \( \gamma_{ij} \) is the proportion of agreement between \( y_i \) and \( y_j \). Then \( \gamma_{ii} = 1, \Gamma \geq 0 \), and therefore

\[
\Delta_2 = \left[ \delta_{ij}^2 \right] = 2 \left[ 1 - \gamma_{ij} \right] = \kappa(\Gamma)
\]

is EDM-2. Hence, all of the information contained in \( \Gamma \) can be embedded in (at most) \( q \)-dimensional Euclidean space, and applying CMDS to \( \Delta = [\delta_{ij}] \) produces the \( d \)-dimensional principal component representation of the complete embedding.

In the case of our \( n = 4 \) film buffs, the matrix \( \Delta_2 \) is EDM-2, but it is not EDM-1.
Conclusions

• Similarities are difficult to interpret unless we assume unit self-similarity, i.e., \( \gamma_{ii} = 1 \). With this assumption, similarities can be interpreted as inner products of points in the positive orthant of the unit sphere.

• If \( B \geq 0 \) with \( b_{ii} = 1 \), then \( b_{ij} \mapsto 2(1 - b_{ij}) \) converts Euclidean inner products to squared Euclidean distances. Hence, \( \gamma_{ij} \mapsto 1 - \gamma_{ij} \) should be interpreted as a transformation from similarity to squared dissimilarity.

• If \( \Gamma \geq 0 \), then \( \Delta_2 = [\delta_{ij}^2] = [1 - \gamma_{ij}] \) is necessarily EDM-2 but not necessarily EDM-1. In this case, \( \Delta = [\delta_{ij}] \) (but not necessarily \( \Delta_2 \)) can be embedded with no loss of information.

• The natural transformation from cosine similarity to dissimilarity is \( \gamma_{ij} \mapsto \arccos \gamma_{ij} \) (angle dissimilarity).
A corpus of text documents is modeled by Latent Dirichlet Allocation, “a generative model that... posits that each document is a mixture of a small number of topics and that each word’s creation is attributable to one of the document’s topics.” LDA associates a probability vector with each document. Based on these probability vectors, you desire to cluster documents. To do so, you require a measure of proximity. What do you do? **What do you do?**