NASA/CR-2002-212136
ICASE Interim Report No. 43

A Dissimilarity Measure for Clustering High- and Infinite Dimensional Data that Satisfies the Triangle Inequality

Eduardo A. Socolovsky
Hampton University, Hampton, Virginia

December 2002

The NASA STI Program Office . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA Scientific and Technical Information (STI) Program Office plays a key part in helping NASA maintain this important role.

The NASA STI Program Office is operated by Langley Research Center, the lead center for NASA's scientific and technical information. The NASA STI Program Office provides access to the NASA STI Database, the largest collection of aeronautical and space science STI in the world. The Program Office is also NASA's institutional mechanism for disseminating the results of its research and development activities. These results are published by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA's counterpart of peer-reviewed formal professional papers, but having less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATIONS. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or cosponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services that complement the STI Program Office's diverse offerings include creating custom thesauri, building customized data bases, organizing and publishing research results . . . even providing videos.

For more information about the NASA STI Program Office, see the following:

- Access the NASA STI Program Home Page at ***http://www.sti.nasa.gov***

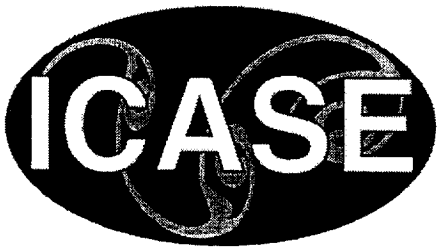
- Email your question via the Internet to help@sti.nasa.gov

- Fax your question to the NASA STI Help Desk at (301) 621-0134

- Telephone the NASA STI Help Desk at (301) 621-0390

- Write to:
  NASA STI Help Desk
  NASA Center for AeroSpace Information
  7121 Standard Drive
  Hanover, MD 21076-1320

NASA/CR-2002-212136
ICASE Interim Report No. 43



A Dissimilarity Measure for Clustering High- and Infinite Dimensional Data that Satisfies the Triangle Inequality

Eduardo A. Socolovsky
Hampton University, Hampton, Virginia

ICASE
NASA Langley Research Center
Hampton, Virginia

Operated by Universities Space Research Association

Prepared for Langley Research Center
under Contract NAS1-97046

December 2002

Available from the following:

NASA Center for AeroSpace Information (CASI)
7121 Standard Drive
Hanover, MD 21076-1320
(301) 621-0390

National Technical Information Service (NTIS)
5285 Port Royal Road
Springfield, VA 22161-2171
(703) 487-4650

A DISSIMILARITY MEASURE FOR CLUSTERING HIGH- AND INFINITE DIMENSIONAL DATA THAT SATISFIES THE TRIANGLE INEQUALITY

EDUARDO A. SOCOLOVSKY[1]

Abstract. The cosine or correlation measures of similarity used to cluster high dimensional data are interpreted as projections, and the orthogonal components are used to define a complementary dissimilarity measure to form a similarity-dissimilarity measure pair. Using a geometrical approach, a number of properties of this pair is established. This approach is also extended to general inner-product spaces of any dimension. These properties include the triangle inequality for the defined dissimilarity measure, error estimates for the triangle inequality and bounds on both measures that can be obtained with a few floating-point operations from previously computed values of the measures. The bounds and error estimates for the similarity and dissimilarity measures can be used to reduce the computational complexity of clustering algorithms and enhance their scalability, and the triangle inequality allows the design of clustering algorithms for high dimensional distributed data.

Subject classification. Applied and Numerical Math

Key words. similarity measures, clustering, high dimensional data, distributed knowledge discovery, scalable data mining

1. Introduction. Clustering is a data analysis technique in which a measure of similarity, or equivalently a measure of dissimilarity, is used to detect groups or patterns in data. Traditionally, these similarity and dissimilarity measures have been related linearly [11]. Clustering of multidimensional data is one of the main tools in Knowledge Discovery from Data (KDD), a field that emerged from the need to extract useful information from the vast amount of data generated by simulations or measurements. Clustering is an essential step in data mining, statistical data analysis, pattern recognition, image processing, and can be used to drive data layout in massive distributed datasets, for example, to improve the retrieval of data subsets from tertiary systems or minimize the amount of data transferred and stored.

The most often used measure of similarity is the Euclidean distance between the vectors representing the data features. This is adequate for low dimensional data, however, for high dimensional data it is well known that the Euclidean distance does not work well. Clustering high-dimensional data in pattern recognition and text and scientific data mining continues to attract a significant amount of attention and effort, since algorithms have to overcome the "dimensionality curse" [8], and simultaneously be scalable and computationally efficient. It has been determined that for high-dimensional data, more adequate measures of similarity are the cosine or Pearson's correlation measure, e.g. see [1-7, 11-13, 16, 17, 21].

The cosine (correlation) similarity measure is the dot-product $U^* V$, where U and V are two unit length (zero mean) vectors representing data features. An important problem with these and other similarity measures in high dimensions is that, *the triangle inequality doesn't hold*, [4, 5]. To illustrate, consider the points $U=(1,1,1,1,1,0,0,0,0,0)/\sqrt{5}$, $V=(1,0,1,0,1,0,1,0,1,0)/\sqrt{5}$, $W=(0,0,0,0,0,1,1,1,1,1)/\sqrt{5}$, then $U^* V = 3/5$, $U^* W = 0$ and $W^* V = 2/5$ which shows that $U^* V \leq U^* W + W^* V$ does not hold.

[1] ICASE and the Data Analysis and Imaging Branch, NASA Langley Research Center, Hampton, VA 23681, while on sabbatical leave from Hampton University. Presently at Norfolk State University. This research was supported by the National Aeronautics and Space Administration under NASA Contract No. NAS1-97046 while the author was in residence at ICASE, NASA Langley Research Center, Hampton, VA 23681-2199.

In this paper, to build upon the suitability of the cosine (correlation) similarity measure for high dimensional data, a non-linearly associated similarity-dissimilarity measure pair is obtained by interpreting the cosine (correlation) similarity measure as the projection of a data point, and defining the associated dissimilarity measure d(V,U) to be the length of the orthogonal component.

It has also been observed that a significant portion of the presently used data analysis techniques become unfeasible for very modest size data sets [19, 20], hence it is important to produce new algorithms, approaches and tools that help extend the limits of computational feasibility and reduce the cost of performing data mining. The key factors are computational efficiency and scalability of the algorithms, as well as the scalability of the implementation. The results presented in this paper are a continuation of [14], and it is expected that they can help enhance the scalability and computational efficiency of algorithms that require a similarity matrix, or multi-step and pre-clustering methods. For instance, algorithms with a canopy or k-means approach could be re-designed to could be re-designed to (at least in the average case) compute only $O(N)$ inner products to generate approximate similarity and dissimilarity matrices or equivalent bounds, instead of the standard $O(N^2)$ inner products.

Distributed hierarchical algorithms to cluster distributed data, construct an approximate global dendrogram from the local dendrograms of the distributed data sets. Generally, they rely on the Euclidean distance using bounds and the triangle inequality. A new hierarchical algorithm for heterogeneously distributed data sets containing data of high dimensions has been designed using a new measure of dissimilarity for distributed data based on the measures studied in this paper. Work is in progress on its implementation and an algorithm for homogeneously distributed data. The algorithms and their results will be reported in a forthcoming paper.

To the best of our knowledge, infinite dimensional clustering is presently not used, but it can potentially be used to cluster results from simulations or observations of phenomena modeled by PDE's whose solutions are an inner product space, e.g., the standard Sobolev spaces. In this case, the data "points" could for example be whole domain finite element simulations or observations at a fixed time, carrying information on the solution and its derivatives.

2. Similarity-Dissimilarity Measures Properties. The *dissimilarity measure d(V,U)* between any two normalized vectors U and V is defined as their "orthogonal distance", i.e., the length of the component of V orthogonal to U. It will be shown that the dissimilarity measure satisfies the symmetric property and the triangle inequality, which yield the standard bound on differences used in metric spaces. As a result of these definitions and properties, bounds on the measures V^*W and d(V,W) between any two normalized vectors V and W are obtained in terms of already computed measures d(V,U), d(W,U), U^*V and U^*W, as shown in the following paragraphs. Specifically, since

$$V = (U^*V)\,U + (I - UU^*)V$$

we define the *dissimilarity measure d(V,U)* by

$$d(V,U) = \| H(I - UU^*)V \|$$

and interchanging the roles of U and V

$$d(U,V) = \| H(I - VV^*)U \|$$

where H is any orthogonal transformation. In this paper $H = I$ the identity is the preferred choice, but for algorithms H could for example be a Householder or Givens transformation.

Property 2.1. $d(V,U) = d(U,V)$

Proof. Using that $V^* V = U^* U = 1$ and the fact that $(I - UU^*)$ and $(I - VV^*)$ are symmetric projections, squaring the dissimilarity measures we obtain:

$$d^2(V,U) = [H(I - UU^*)V]^* [H(I - UU^*)V] = V^*(I - UU^*)V = 1 - (U^* V)^2$$

$$d^2(U,V) = [H(I - VV^*)U]^* [H(I - VV^*)U] = U^*(I - VV^*)U = 1 - (U^* V)^2 \quad \blacksquare$$

Property 2.2. Given arbitrary unitary vectors U, V and W

$$d(V,W) \leq d(V,U) + d(U,W)$$

Proof. Let P be the projection of V in the direction of U

(1) $$P = (U^*V)U$$

then $V - P = V - (U^*V)U = (I - UU^*)V$ is the component of V orthogonal to U, consequently

(2) $$d(V,U) = \| V - P \|$$

Similarly, if $S = (W^*U)W$ is the projection of U in the direction of W, then $U - S = (I - WW^*)U$ is the component of U orthogonal to W, and

(3) $$d(U,W) = \| U - S \|$$

Now, let $Q = (W^*P)W$ be the projection of P in the direction of W, then from (1) and (3)

(4) $$\|P - Q\| = \| (I - WW^*)P\| = |U^*V| \, \| (I - WW^*)U\| \leq \|U - S\| = d(U,W)$$

Also, let $R = (W^*V)W$ be the projection of V in the direction of W and $V - R = (I - WW^*)V$ be the component of V orthogonal to W, then $\| V - R \| \leq \| V - Q \|$ and $d(V,W) = \| V - R \|$, which yields

(5) $$d(V,W) \leq \| V - Q \|$$

Finally, from (5), (2) and (4)

$$d(V,W) \leq \| V - P \| + \|P - Q\| \leq d(V,U) + d(U,W) \quad \blacksquare$$

Notice that the dissimilarity measure *is not a distance*, since $d(V,U) = 0$ when either $U = V$ or $U = -V$. A direct consequence of the definition of $d(V,U)$ and Properties 2.1 and 2.2 are:

Property 2.3. $U^* V > \delta$ if and only if $d^2(V,U) < 1 - \delta^2$ and $U^* V > 0$ $\blacksquare$

Property 2.4. $d(V,W) \geq | d(V,U) - d(W,U) |$ $\blacksquare$

From properties 2.3 and 2.4 immediately follows

Property 2.5. If d(V,U) and d(W,U) have been previously computed and
$| d(V,U) - d(W,U) | \geq \sqrt{1-\delta^2}$, then $W^* V \leq \delta$. ■

If the basic clustering criterion is summarized by: *"W and V are in the same cluster if and only if* $W^*V > \delta$, *with* δ *near 1"*, Property 2.3 yields an equivalent statement in terms of d(W,V) and Property 2.5 shows that if the difference of the dissimilarities d(V,U) and d(W,U) is large enough, then the vectors V and W cannot be in the same cluster. Next, two equivalent bounds on W^* V are given in terms of previously computed similarity and dissimilarity measures

Property 2.6. $V^*W \leq (U^*V)(U^*W) + d(V,U)\, d(W,U)$

Proof. From the orthogonal factorizations V = P+ (V - P) and W = T + (W – T), and Cauchy-Schwartz we have

$$V^*W \leq (U^*V)(U^*W) + \|V - P\| \, \|W - T\|$$

the definition of the dissimilarity measure yields the result. ■

Property 2.7. $$V^*W \leq 1 - \frac{\left(U^*W - U^*V\right)^2}{2} - \frac{\left|d(V,U) - d(W,U)\right|^2}{2}$$

Proof. From

(6) $$V^*W = \frac{1}{2}\left[\|W\|^2 + \|V\|^2 - \|W - V\|^2\right] = 1 - \frac{\|W - V\|^2}{2}$$

and the orthogonal decomposition of W – V

$$W - V = (W - T) - (V - P) + \left(U^*W - U^*V\right)U$$

where P is given by (1) and T = (U^* W)U. By Pythagora's theorem

(7) $$\|W - V\|^2 = \|(W - T) - (V - P)\|^2 + \left(U^*W - U^*V\right)^2$$

and for any orthogonal transformation H

$$\|(W - T) - (V - P)\| \geq \|H(W - T)\| - \|H(V - P)\|$$

and substituting in (7), we obtain

(8) $$\|W - V\|^2 \geq \|H(W - T)\| - \|H(V - P)\|^2 + \left(U^*W - U^*V\right)^2$$

substituting (8) into (6), using $d(V,U) = \|H(V - P)\|$ and $d(W,U) = \|H(W - T)\|$ yields the result. ■

Clearly, since there is a whole hyperplane orthogonal to U, properties 2.4 and 2.6-7 won't provide sharp or conclusive bounds in cases in which the orthogonal lengths dominate and their difference is not large enough. However, properties 2.4 and 2.6-7 can be complementary in other cases. Property 2.4 is inconclusive when the right side is small, i.e., $| d(V,U) - d(W,U) | = \varepsilon$ with $\varepsilon < \sqrt{1-\delta^2}$, since it only says $d(V,W) \geq \varepsilon$, however property 2.7 yields

$$V^*W \le 1 - \frac{(U^*W - U^*V)^2}{2} - \frac{\varepsilon^2}{2}$$

and for $(U^*W - U^*V)^2$ sufficiently large we obtain $V^*W \le \delta$, i.e., W and V are not in the same cluster. For example, if the components of V and W orthogonal to U are equal (i.e., $\varepsilon = 0$), and the components of V and W in the direction of U are of equal length but opposite sign (i.e., $U^*V = -U^*W$) and $\frac{1-\delta}{2} \le (U^*V)^2$ we obtain $V^*W \le 1 - 2(U^*V)^2 \le \delta$. Conversely, property 2.5 may hold while property 2.7 is inconclusive when $|U^*V - U^*W| < 1 - \delta$. In this case, from 2.7 it can't be concluded that $V^*W \le \delta$ since $1 + \delta^2 - |U^*V - U^*W|^2 > 2\delta$, and the right side of 2.7 is less than or equal to $1 - \frac{1-\delta^2}{2} - \frac{|U^*V - U^*W|^2}{2}$.

3. The Infinite Dimensional Case. The properties of section 2 also hold in a general inner-product space **H**. The proofs share the ideas of the finite dimensional case but require a different formalism, which is briefly presented in this section. Given an unitary vector U we define a map $\Phi_U : \mathbf{H} \to \mathbf{H}$, by

$$\Phi_U(V) = V - \langle U, V \rangle U \qquad \text{for any } V \text{ in } \mathbf{H}$$

The *dissimilarity measure d(V,U) between any two normalized vectors U and V* is now defined by

$$d(V,U) = \|\Phi_U(V)\|$$

Notice that $\Phi_U(V)$ is the orthogonal component of V with respect to U, since for any V in **H**

$$V = \langle U, V \rangle U + \Phi_U(V) \tag{9}$$

and for any α

$$\langle \alpha U, \Phi_U(V) \rangle = \alpha \langle U, V - \langle U, V \rangle U \rangle = 0 \tag{10}$$

consequently, for any β

$$\|V - \beta U\|^2 = \|\Phi_U(V) + (\langle U, V \rangle - \beta)U\|^2 = \|\Phi_U(V)\|^2 + |\langle U, V \rangle - \beta|^2 \tag{11}$$

The map Φ_U has similar properties to the matrix (I – UU*) for the finite dimensional case

Property 3.1. Φ_U is a self-adjoint projection.

Proof. For any V and W in **H**, by (9) and (10)

$$\langle W, \Phi_U(V) \rangle = \langle \Phi_U(W), \Phi_U(V) \rangle + \langle \Phi_U(W), \langle U, V \rangle U \rangle = \langle \Phi_U(W), V \rangle$$

$$\Phi_U^2(V) = \Phi_U(V) - \langle U, \Phi_U(V) \rangle U = \Phi_U(V) \quad \blacksquare$$

The dissimilarity measure has the same properties as in the finite dimensional case. In effect, for any unit vectors U, V and W, we have

Property 3.2. $d(V,U) = d(U,V)$

Proof. Using Property 3.1 and $<U,U>=<V,V>=1$, squaring the dissimilarity measures we obtain:

$$d^2(U,V) = <\Phi_U(V), \Phi_U(V)> = <\Phi_U(V), V> = 1 - <U,V>^2$$

$$d^2(V,U) = <\Phi_V(U), \Phi_V(U)> = <\Phi_V(U), U> = 1 - <V,U>^2 \quad \blacksquare$$

Property 3.3. $d(V,W) \le d(V,U) + d(U,W)$

Proof. Let $P = <U,V> U$ be the projection of V in the direction of U, then from (9) $V - P = \Phi_U(V)$, and

$$d(V,U) = \|V - P\| \tag{12}$$

Let $Q = <W,P> W$ be the projection of P in the direction of W, then from (9)

$$\|P - Q\| = |<U,V>| \, \|U - <W,U> W\| \le \|\Phi_W(U)\| = d(W,U) \tag{13}$$

Finally, let $R = <W,V> W$ be the projection of V in the direction of W, then from (9), (11), (12) and (13)

$$d(V,W) = \|V - R\| \le \|V - Q\| \le \|V - P\| + \|P - Q\| \le d(V,U) + d(U,W) \quad \blacksquare$$

Substituting dot product notation by inner product notation, it is straightforward to verify that the rest of the properties in section 2 also hold in a general inner-product space **H**.

4. Error Estimates for the Triangle Inequality. The error introduced in approximating d(W,X) by d(W,Y)+d(Y,X), $W \neq X$, is discussed in this section. The first result obtained is confirmation of the intuitive idea that the "raise" of Y, i.e. the distance of Y to span(W,X), is one of the two added independent components of the error. Then, estimates for the other component of the error are obtained by considering Y in span(W,X).

Let Y' be the projection of Y onto span(W,X), i.e., $\langle Y - Y', W \rangle = 0$ and $\langle Y - Y', X \rangle = 0$, then

$$d^2(W,Y) = \|Y - \langle Y,W\rangle W\|^2 = \|Y - Y'\|^2 + \|Y' - \langle Y',W\rangle W\|^2 = \|Y - Y'\|^2 + \|Y'\|^2 d^2(W,Y^*)$$

$$d^2(Y,X) = \|Y - \langle Y,X\rangle X\|^2 = \|Y - Y'\|^2 + \|Y' - \langle Y',X\rangle X\|^2 = \|Y - Y'\|^2 + \|Y'\|^2 d^2(Y^*,X)$$

where $Y^* = \dfrac{Y'}{\|Y'\|}$ and since $\|Y'\|^2 = 1 - \|Y - Y'\|^2$, it follows that

$$d^2(Y,X) = \|Y - Y'\|^2 (1 - d^2(Y^*,X)) + d^2(Y^*,X)$$

$$d^2(W,Y) = \|Y - Y'\|^2 (1 - d^2(W,Y^*)) + d^2(W,Y^*)$$

which shows that

$$d(W,Y) + d(Y,X) \geq d(W,Y^*) + d(Y^*,X)$$

Consequently

$$\inf_Y d(W,Y) + d(Y,X) = \inf_{Y \in span(W,X)} d(W,Y) + d(Y,X)$$

and to find error bounds and estimates, the above justifies focusing on

(14) $$Y = \alpha W + \beta X, \qquad \|Y\| = 1$$

From the definition of $d(\cdot,\cdot)$, using (14) and the fact that X and W are unit length, it follows that

(15) $$\begin{aligned} d(W,Y) + d(Y,X) &= \|\alpha W + \beta X - \langle \alpha W + \beta X, W\rangle W\| + \|\alpha W + \beta X - \langle \alpha W + \beta X, X\rangle X\| = \\ &= |\beta| \|X - \langle X, W\rangle W\| + |\alpha| \|W - \langle W, X\rangle X\| = \\ &= (|\alpha| + |\beta|)\, d(W,X) \end{aligned}$$

In summary, the fundamental result of this section obtained from the triangle inequality and (15), can be stated as

Property 4.1. For Y in the span(W,X)

(16) $$d(W,X) \leq d(W,Y) + d(Y,X) = (|\alpha| + |\beta|)\, d(W,X)$$ ■

Motivated by Property 4.1, the rest of this section concentrates on obtaining bounds for $(|\alpha| + |\beta|)$. From (14)

(17) $$1 = \alpha^2 + \beta^2 + 2\alpha\beta\langle W,X\rangle$$

and it follows from (17) that

(18) $$(|\alpha| + |\beta|)^2 = 1 + 2|\alpha|\,|\beta| - 2\alpha\beta\langle W,X\rangle$$

which shows that $|\alpha| + |\beta| > 1$, and that for the Y that yields a minimum of $|\alpha| + |\beta|$ it is necessary that

(19) $$\alpha\beta\langle W,X\rangle \geq 0.$$

Consequently, (18) can be rewritten

(20) $$|\alpha| + |\beta| = \sqrt{1 + 2|\alpha|\,|\beta|\,(1 - |\langle W,X\rangle|)}$$

On the other hand, for any constant c, $0 < c < 1$, from (17) it follows

$$\alpha^2 < \{1 - 2\alpha\beta\langle W,X\rangle\} - c\beta^2 \quad \text{and} \quad \beta^2 < \{1 - 2\alpha\beta\langle W,X\rangle\} - c\alpha^2$$

multiplying

$$\alpha^2\beta^2 < \{1-2\alpha\beta\langle W,X\rangle\}\left|(1-c+c)\{1-2\alpha\beta\langle W,X\rangle\}-c\alpha^2-c\beta^2\right|+c^2\alpha^2\beta^2$$

and using (17)

$$\left(1-c^2\right)\alpha^2\beta^2 < (1-c)\{1-2\alpha\beta\langle W,X\rangle\}^2$$

which yields

$$|\alpha||\beta| < \frac{1}{\sqrt{1+c}}\{1-2\alpha\beta\langle W,X\rangle\}$$

defining $k=\frac{1}{\sqrt{1+c}}$ and using (19) to introduce absolute values

$$\{1+2k|\langle W,X\rangle|\}|\alpha||\beta| < k$$

which gives

$$|\alpha||\beta| < \frac{1}{\sqrt{1+c}+2|\langle W,X\rangle|}$$

and taking the limit as $c\to 1$

(21)
$$|\alpha||\beta| \le \frac{1}{\sqrt{2}+2|\langle W,X\rangle|}$$

Finally substituting (21) in (20), the following bound is obtained

Property 4.2. For any Y in span(W,X) satisfying (14) and (19)

$$|\alpha|+|\beta| \le \sqrt{1+\frac{2}{\sqrt{2}+2|\langle W,X\rangle|}\left(1-|\langle W,X\rangle|\right)} \quad \blacksquare$$

Table 1 was obtained from Property 4.2, and lists bounds for $|\alpha|+|\beta|$ for some standard values of $|\langle W,X\rangle|$ (or equivalently, cosine of angles between W and X):

TABLE 1

For $\lvert\langle W,X\rangle\rvert \ge$	$1/2$	$\sqrt{2}/2$	$\sqrt{3}/2$
$\lvert\alpha\rvert+\lvert\beta\rvert \le$	$\sqrt[4]{2} \approx 1.1892$	$\sqrt{(1+\sqrt{2})/2} \approx 1.09868$	$\sqrt{1+(2-\sqrt{3})(\sqrt{3}-\sqrt{2})} \approx 1.0417$

5. Optimal Error Estimates for the Triangle Inequality. In this section conditions to minimize $d(W,Y)+d(Y,X)$ are sought. First the optimal directions are determined and then the error for those

directions is found. The arguments given in section 4 show that it is sufficient to consider the set of Y's satisfying (14), and according to (16) the optimum is found minimizing $|\alpha| + |\beta|$. For convenience Y is written

(22) $$Y = \frac{a W + b X}{\| a W + b X \|} = \alpha W + \beta X$$

so that $|\alpha| + |\beta| = f(a,b)$ becomes a function of a and b given by

(23) $$f(a,b) = \frac{|a| + |b|}{\sqrt{a^2 + 2ab\langle W, X\rangle + b^2}}$$

A standard gradient calculation shows that $\frac{\partial f}{\partial a} = 0$ if and only if

(24) $$sign(a)(a^2 + 2ab\langle W, X\rangle + b^2) - (|a| + |b|)(a + b\langle W, X\rangle) = 0$$

and (24) reduces to

(25) $$(|a| - |b|)\,|b|\,(sign(b)\langle W, X\rangle - sign(a)) = 0$$

Equation (25) holds for

(26) $$|a| = |b|$$

or for $\langle W, X\rangle = sign(a)/sign(b)$, which is the trivial case where W is co-linear with X. Similar results are obtained from $\frac{\partial f}{\partial b} = 0$. From (19) and (26) it follows that

(27) $$a = b \qquad for \quad \langle W, X\rangle > 0$$

(28) $$a = -b \qquad for \quad \langle W, X\rangle < 0$$

Substituting in (23), both (27) and (28) yield

(29) $$f(a,b) = \sqrt{\frac{2}{1 + |\langle W, X\rangle|}}$$

Table 2 was obtained from (29) and lists *optimal* bounds for $|\alpha| + |\beta|$ for some standard values of $|\langle W, X\rangle|$ (or equivalently, cosine of angles between W and X):

TABLE 2

For $\lvert\langle W, X\rangle\rvert \geq$	$1/2$	$\sqrt{2}/2$	$\sqrt{3}/2$
$\lvert\alpha\rvert + \lvert\beta\rvert \leq$	$2/\sqrt{3} \approx 1.1547$	$\sqrt{2}\sqrt{(2-\sqrt{2})} \approx 1.08239$	$2/\sqrt{2+\sqrt{3}} \approx 1.03528$

TABLE 3

$\langle W, X\rangle$	0	$1/2$	$\sqrt{2}/2$	$\sqrt{3}/2$
$d(W,X)$	1	$\sqrt{3}/2$	$\sqrt{2}/2$	$1/2$
$d(W,Y) + d(Y,X)$	$\sqrt{2}$	1	$\sqrt{2-\sqrt{2}} \approx 0.765367$	$2\sqrt{2-\sqrt{3}} \approx 0.517638$

Table 3 illustrates the estimates obtained with the triangle inequality for the optimal case (27) and some standard values of $\langle W, X \rangle$

REFERENCES

[1] A. BANERJEE AND J. GOSH, *On Scaling Balanced Clustering Algorithms*, Proceedings of the Second SIAM International Conference on Data Mining, R. Grossman *et al.* (eds.), 2002, ISBN 0-89871-517-2.

[2] I.B. CRABTREE AND S. SOLTYSIAK, *Identifying and Tracking Changing Interests*, International Journal of Digital Libraries, 2 (1998), pp. 38-53.

[3] I. DHILLON, Y. GUAN AND J. KOGAN, *Refining Clusters in High Dimensional Text Data*, Proceedings of the Workshop on Clustering High Dimensional Data and its Applications, Second SIAM International Conference on Data Mining, (2002), pp. 71-82.

[4] L. ERTOZ, M. STEINBACH AND V. KUMAR, *A New Shared Nearest Neighbor Clustering Algorithm and its Applications*, Proceedings of the Workshop on Clustering High Dimensional Data and its Applications, Second SIAM International Conference on Data Mining, (2002), pp 105-116.

[5] L. ERTOZ, M. STEINBACH AND V. KUMAR: *Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data*, CS Technical Report, Univ. of Minnesota, (2002), available as a PDF file at http://www-users.cs.umn.edu/~kumar/papers/papers.html#bbbb

[6] L. FONER, *A Multi-Agent Referral System for Matchmaking*, Proceedings of the First International Conference on the Practical Applications of Agents and Multi-Agent Systems, B. Crabtree and N. Jennings (eds.), The Practical Application Company, London, UK, pp. 245-262, 1996.

[7] L. FONER, *Political Artifacts and Personal Privacy: The Yenta Multi-Agent Distributed Matchmaking System*, PhD Dissertation, Media Arts and Sciences Program, Massachusetts Institute of Technology, June 1999.

[8] A. HINNENBURG AND D. KEIM, *Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering*, VLDB'99, Morgan Kaufman, (1999), pp. 506-517, ISBN 1-55860-615-7.

[9] E. JOHNSON AND H. KARGUPTA, *Collective, Hierarchical Clustering from Distributed Heterogeneous Data*, Lecture Notes in Computer Science 1759, Springer Verlag, 2000.

[10] H. KARGUPTA, W. HUANG, K. SIVAKUMAR AND E. JOHNSON, *Distributed Clustering Using Collective Principal Component Analysis*, to appear in Knowledge and Information Systems.

[11] L. KAUFMAN AND P.J. ROUSSEEUW, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons Inc., New York, 1990.

[12] V. KUMAR , M. STEINBACH, P. TAN, S. KLOOSTER, C. POTTER, AND A. TORREGROSA, *Mining Scientific Data: Discovery of Patterns in the Global Climate System*, 2001 Joint Statistical Meeting.

[13] G. PROCTOR AND C. WINTER, *Information Flocking: Data Visualization in Virtual Worlds Using Emergent Behaviours*, Proceedings of the First International Conference on Virtual Worlds, J.-C. Heudin (ed.), Springer-Verlag, pp. 168-176, 1998.

[14] K. SEVERANCE, P. BREWSTER, D. CORDNER, K. LODDING, D. NARK, AND E. SOCOLOVSKY, *Data Swarming: A Metaphor for Discovery of Emergent Patterns among Qualitative and Quantitative Data*, NASA LaRC Creativity and Innovation Grant 2210.726100 Final Report, 2001 Director's Discretionary Fund Report, March 2002, pp. 188.

[15] Workshop on Clustering High Dimensional Data and its Applications, April 13, 2002, Arlington, VA. Held in Conjunction with the Second SIAM International Conference on Data Mining (SDM 2002).

[16] A. STREHL, J. GHOSH AND R. MOONEY, *Impact of Similarity Measures on Web Page Clustering*, AAAI 2000 Workshop of Artificial Intelligence for Web Search, July 2000.

[17] M. STEINBACH, P. TAN, V. KUMAR , S. KLOOSTER, AND C. POTTER, *Data Mining for the Discovery of Ocean Climate Indices*, Proceedings of the Fifth Workshop on Scientific Data Mining, Second SIAM International Conference on Data Mining, (2002), pp. 7-16.

[18] N.F. SAMATOVA, G. OSTROUCHOV, A. GEIST, AND A. MELECHKO, *RACHET: An Efficient Cover-Based Merging of Clustering Hierarchies from Distributed Datasets*, International Journal of Distributed and Parallel Databases, 2001, accepted for publication.

[19] E. WEGMAN, *Huge data sets and the frontiers of computational feasibility*, Journal of Computational and Graphical Statistics, 4(4) (1995), pp. 281-295.

[20] E. WEGMAN, *Visions: New techniques and technologies in statistics*, Computational Statistics, 15 (2000), pp. 133-144.

[21] WU, M.; FULLER, M.; AND WILKINSON, R., *Using Clustering and Classification Approaches in Interactive Retrieval*, Information Processing and Management, 37(3) (2001), pp. 452-484.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December 2002	Contractor Report

4. TITLE AND SUBTITLE
A DISSIMILARITY MEASURE FOR CLUSTERING HIGH- AND INFINITE DIMENSIONAL DATA THAT SATISFIES THE TRIANGLE INEQUALITY

5. FUNDING NUMBERS
C NAS1-97046
WU 505-90-52-01

6. AUTHOR(S)
Eduardo A. Socolovsky

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
ICASE
Mail Stop 132C
NASA Langley Research Center
Hampton, VA 23681-2199

8. PERFORMING ORGANIZATION REPORT NUMBER
ICASE Interim Report No. 43

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Langley Research Center
Hampton, VA 23681-2199

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
NASA/CR-2002-212136
ICASE Interim Report No. 43

11. SUPPLEMENTARY NOTES
Langley Technical Monitor: Dennis M. Bushnell
Final Report

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified–Unlimited
Subject Category 64
Distribution: Nonstandard
Availability: NASA-CASI (301) 621-0390

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*
The cosine or correlation measures of similarity used to cluster high dimensional data are interpreted as projections, and the orthogonal components are used to define a complementary dissimilarity measure to form a similarity-dissimilarity measure pair. Using a geometrical approach, a number of properties of this pair is established. This approach is also extended to general inner-product spaces of any dimension. These properties include the triangle inequality for the defined dissimilarity measure, error estimates for the triangle inequality and bounds on both measures that can be obtained with a few floating- point operations from previously computed values of the measures. The bounds and error estimates for the similarity and dissimilarity measures can be used to reduce the computational complexity of clustering algorithms and enhance their scalability, and the triangle inequality allows the design of clustering algorithms for high dimensional distributed data.

14. SUBJECT TERMS
similarity measures, clustering, high dimensional data, distributed knowledge discovery, scalable data mining

15. NUMBER OF PAGES
16

16. PRICE CODE
A03

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified		