Similarity, kernels, and the triangle inequality

Frank Jäkel
Technische Universität Berlin
FR 6-4, Franklinstr. 28/29
10587 Berlin, Germany

Bernhard Schölkopf
Max Planck Institute for Biological Cybernetics
Spemannstr. 38
72076 Tübingen, Germany

Felix A. Wichmann
Technische Universität Berlin
FR 6-4, Franklinstr. 28/29
10587 Berlin, Germany

Similarity is used as an explanatory construct throughout psychology and multidimensional scaling (MDS) is the most popular way to assess similarity. In MDS similarity is intimately connected to the idea of a geometric representation of stimuli in a perceptual space. Whilst connecting similarity and closeness of stimuli in a geometric representation may be intuitively plausible, Tversky and Gati (1982) have reported data which are inconsistent with the usual geometric representations that are based on segmental additivity. We show that similarity measures based on Shepard's universal law of generalization (Shepard, 1987) lead to an inner product representation in a reproducing kernel Hilbert space. In such a space stimuli are represented by their similarity to all other stimuli. This representation, based on Shepard's law, has a natural metric that does not have additive segments whilst still retaining the intuitive notion of connecting similarity and distance between stimuli. Furthermore, this representation has the psychologically appealing property that the distance between stimuli is bounded.
This is a preprint of an article that appeared in Journal of Mathematical Psychology 52(5) 297-303 (2008). The preprint may differ from the published version.

The most influential approach to model similarity has been geometrical. The central idea in this approach is that stimuli are represented in a perceptual space and the distance between stimuli in this space determines their similarity. In the simplest case the space is assumed to be Euclidean and the similarity of stimuli decreases with their distance in space. Multidimensional scaling (MDS) provides a class of algorithms that make it possible to reconstruct the coordinates in the putative perceptual space from similarity data, for example similarity ratings or confusion probabilities. Shepard (1987) argued that the best experimental measure for similarity are generalization gradients. He further presented data that indicated that generalization gradients are an exponential function of the distance in perceptual space. This relationship between generalization gradients and perceptual spaces is often referred to as Shepard's *universal law of generalization*.

In a well-known series of papers Tversky and colleagues have challenged the idea of a geometric representation (Beals, Krantz, & Tversky, 1968; Tversky, 1977; Tversky & Gati, 1982). They provided convincing evidence that geometric representations cannot account for many human similarity judgments. Even though their criticism has been substantial, MDS has been used in practice with considerable success. Categorization models in particular have relied heavily on geometric representations—seemingly unfazed by Tversky's criticism (Nosofsky, 1986). In this note we will reconcile Tversky's critique with Shepard's universal law of generalization. Read carefully, Tversky's most fundamental critique does not exclude the possibility of a geometric perceptual space per se, it only attacks the commonly used metrics with additive segments (Tversky & Gati, 1982). This class, however, includes many intuitive geometries: Euclidean spaces, spaces with a Minkowski p-norm, and curved Riemannian geometries. We will introduce a representation of the perceptual space that arises naturally from Shepard's law and that is not affected by Tversky's criticism. This representation has several psychologically interesting properties: It does not have additive segments, it is bounded and it represents stimuli by their similarity to all other stimuli (Edelman, 1998). Furthermore, it provides a deeper understanding of the constraints that Shepard's law imposes on data and on their embedding in a psychological space. The representation that we suggest is based on the mathematical theory of reproducing kernel Hilbert spaces that can be used to model the similarity of stimuli as inner products.

Similarity

There have been early attempts to model similarity judgments as inner products. Ekman (1954) made the assumption that stimuli are represented in the mind as vectors in a multidimensional Euclidean space and that the similarity of points is given by their inner product (Gregson, 1975; Borg & Groenen, 1997, both provide an overview on Ekman's approach). A little earlier, Torgerson (1952) presented

a method that is now widely known as classical multidimensional scaling. Instead of requiring a direct measurement of similarity this method indirectly determined the dissimilarity between stimuli by using the method of triads. Under the assumption that dissimilarity is linear with distance in a Euclidean space it is possible to reconstruct the coordinates of the stimuli in a perceptual space using a procedure that was suggested by Young and Householder (1938).

There was no a priori reason to believe that mental representations should be Euclidean. There was ever little reason to believe that measurements of similarity or dissimilarity were linearly related to the distance or the inner product in a Euclidean space. In search for an alternative, Shepard (1962) and Kruskal (1964) developed ordinal multidimensional scaling methods that allow for arbitrary metrics and for a non-linear relationship between distances in a metric space and a so-called proximity measure. By proximity measure they referred to both, similarity or dissimilarity measurements. The key idea is that a dissimilarity measure has to be monotonically increasing with the distance in the metric space and a similarity measure has to be monotonically decreasing with the distance in the metric space. Their algorithms only use ordinal properties of the data and allow the convenient use of indirect measurements like confusion probabilities and reaction times as a proximity measure. Today, the use of proximity measures that are monotonically related to a metric is the prevailing approach and the idea that similarity could be directly modeled as an inner product has seemingly vanished. We argue, however, that inner products still deserve a place in theories of similarity. In fact, we will show that inner products in the form of so-called positive definite kernels have been used extensively without being recognized as inner products. In order to do so some common assumptions about mental representations and psychological distance will be made explicit in the following.

Psychological space and distance

It is tempting to assume that the representation of a stimulus is given as a point in a vector space. The dimensions of the vector space ought to describe the perceptual dimensions along which stimuli can vary. With respect to the norm in this space it has become customary to use a weighted ℓ_p norm for the length of a n-dimensional vector x:

$$\|x\|_p = \left(\sum_{i=1}^{n} \alpha_i \, |x_i|^p \right)^{\frac{1}{p}}. \tag{1}$$

with positive weights α_i. The weights are needed to allow for systematic variations of the norm over tasks or over individuals. This norm induces a metric on the space (which is also known as the Minkowski p-metric or power model):

$$d_p(x, y) = \|x - y\|_p = \left(\sum_{i=1}^{n} \alpha_i \, |x_i - y_i|^p \right)^{\frac{1}{p}}. \tag{2}$$

On a first glance, this metric seems to be an ad-hoc choice but it is implied by a set of desirable axioms that include the metric axioms, segmental additivity and conditions on the dimensions and the combination of dimensions (Tversky & Krantz, 1970).

The ℓ_p formula is a norm and induces a metric only for $p \geq 1$. For $p < 1$ equation (2) does not fulfill the triangle inequality—an issue that is crucial for psychology and that will be discussed below. Irrespective of whether d_p is a metric or not we will call it a distance (Blumenthal, 1953). We will only call it a metric if it also satisfies the triangle inequality (i.e. for $p \geq 1$). A long list of studies used the ℓ_p norm either directly or in the form of the Euclidean or city-block metric, that is with $p = 2$ or $p = 1$, respectively (Attneave, 1950; Shepard, 1964; Garner, 1974; Nosofsky, 1986; Kruschke, 1992; Love, Medin, & Gureckis, 2004).

Generalization gradients

If one is willing to commit oneself to a vectorial representation of stimuli and the distance d_p on this space there is still the question of how the distance in this space relates to the measured (dis)similarity of the stimuli. Intuitively, similarity should decrease and dissimilarity increase with distance.

Shepard (1987) argued that the best measure for similarity are generalization gradients. He analyzed several data-sets with his ordinal multidimensional scaling method and found that the non-linear relationship between the distance in the psychological ℓ_p space and the measured similarity is generally monotonic and, in Shepard's terms, concave upward. In its stronger version Shepard's claim is that the relationship is exponentially decreasing. We refer to this exponential relationship as the universal law of generalization. Shepard's finding was in accordance with his much earlier suggestion of the exponential as a link between confusion probabilities and psychological distance (Shepard, 1957) and his diffusion model of similarity (Shepard, 1958). Furthermore, he tried to deduce the exponential from assumptions on optimal classification performance (Shepard, 1987; Tenenbaum & Griffiths, 2001; Chater & Vitanyi, 2003). Shepard's work has been extremely influential and has led others to use the exponential law (Nosofsky, 1986; Kruschke, 1992; Love et al., 2004, e.g.). In this framework, a very general formulation for the similarity between two representations x and y is:

$$k(x, y) = \exp(-d_p(x, y)^q), \tag{3}$$

an exponential of the distance d_p (2) raised to the power of q. Shepard's original formulation did not have the exponent q but other authors make use of this extra parameter (Nosofsky, 1990; Ashby & Maddox, 1993).

Kernels

Under certain circumstances the similarity measure as given by (3) is a so-called positive definite kernel and therefore opens up the rich theory of Hilbert spaces for the analysis of similarity. A complete and possibly infinite dimensional vector space with an inner product is called a Hilbert space. We will describe such a Hilbert space that is associated with Shepard's universal law of generalization but

we will not give a detailed introduction to the mathematics involved—a tutorial introduction that makes connections to the machine learning and neural networks literature can be found in a companion paper (Jäkel, Schölkopf, & Wichmann, 2007). Our brief discussion here follows Schölkopf and Smola (2002), leaving out many of the technical details.

A positive definite similarity function

A real and symmetric function $k(\cdot,\cdot)$ is called a positive definite kernel if for all choices of N points $x_1, ..., x_N$ from the domain of k, the following holds:

$$\sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j k(x_i, x_j) \geq 0 \tag{4}$$

for all possible real coefficients w_i. If k is a psychological similarity function and the x_i are stimuli this means that for all possible stimuli the matrix of pairwise similarities is always positive semi-definite.

The real and symmetric function $k(x, y)$ as given in (3) is such a positive definite kernel only for certain choices of q and p. For the current discussion it is enough to first restrict attention to the simpler case $q = p$:

$$k(x, y) = \exp(-d_p(x, y)^p) = \exp(-\sum_{i=1}^{n} \alpha_i \, |x_i - y_i|^p). \tag{5}$$

The $q = p$ case becomes what Nosofsky (1990) called "interdimensional multiplicative" because similarities are calculated for each dimension and then multiplied. With p chosen to be two the similarity measure has the form of a Gaussian kernel. With p chosen to be one the function is sometimes called Laplacian. These two cases correspond to the Euclidean and the city-block metric, respectively. Figure 1 shows the similarity kernel for $p = 2$, $p = 1$ and $p = \frac{1}{2}$. Contrary to the Gaussian kernel the other two kernels have clearly defined axes.

While the Minkowski p-metric (2) only defines a metric for $p \geq 1$ the similarity measure in (5) is a positive definite kernel for $0 < p \leq 2$. This is a classic result on positive definite kernels (Schoenberg, 1938). We can even give a slightly more general result for the case where q does not equal p (Schoenberg, 1938, using Corollary 2). For $p \leq 2$ Eq. (3) is a positive definite kernel if $0 < q \leq p$. The conditions for $p > 2$ are more complicated but known results are summarized by Koldobsky and Koenig (2001). To the best of our knowledge there is no paper in psychology that claims a value for p bigger than two. Thus, concentrating on the case where $p \leq 2$ appears to be no serious restriction. Note that the most interesting cases of the similarity kernel that have been reported in the literature are all positive definite. The Laplacian kernel ($q = p = 1$) and the Gaussian kernel ($q = p = 2$) are positive definite but also Shepard's original suggestion with $q = 1$ and $p = 2$.

For the rest of the paper we will concentrate on the case $q = p$ and especially on the case where $p < 1$ because there are several reports for a p smaller than one in the literature (Shepard, 1964; Tversky & Gati, 1982; Indow, 1994; Lee, 2008). In these cases trying to model similarity with a Minkowski metric is problematic because (2) is not a metric if $p < 1$—but the axioms of a metric space have been essential in the development of MDS and in the interpretation of results. The similarity measure in (5) is, however, still a positive definite kernel for $0 < p < 1$ and therefore the kernel framework might provide us with an alternative interpretation.

Reproducing kernel Hilbert space

We have observed that the above measure for similarity is a positive definite kernel. We will now introduce a vector space using this positive definite kernel as an inner product. Let us assume, for simplicity, that the perceptual space is $\mathbb{R}^n$. The vector space $\mathcal{H}$ that will be constructed below is a space of real functions defined on the perceptual space, that is a function f in the vector space $\mathcal{H}$ is of the form $f : \mathbb{R}^n \to \mathbb{R}$.

The crucial idea is that we associate each stimulus with its similarity to all other stimuli (Edelman, 1998). For each stimulus x in the perceptual space there is a function from $\mathbb{R}^n$ to $\mathbb{R}$ that captures the similarity of x to all other stimuli in the perceptual space. This function is $k(\cdot, x)$ with a fixed x and interpreted as a function of its first argument. This function lies in the vector space $\mathcal{H}$ that we will construct. In this way, we associate each stimulus x in the perceptual space with a function, its similarity function, in $\mathcal{H}$. Instead of examining the perceptual space directly we will analyze the space of functions $\mathcal{H}$ that is defined on the perceptual space and that contains all the similarity functions associated with each stimulus. It will turn out that this space has psychologically interesting properties. We will denote the function that maps each stimulus to its similarity function in $\mathcal{H}$ with $\Phi : \mathbb{R}^n \to \mathcal{H}$ and define it to be

$$\Phi(x) = k(\cdot, x). \tag{6}$$

The vector space $\mathcal{H}$ is now defined to be the set of functions that can be described as a finite linear combination of similarity functions. Each function f in $\mathcal{H}$, by definition, can be written as

$$f(x) = \sum_{i=1}^{N} w_i k(x, x_i) \tag{7}$$

for some N and a choice of points $x_1, ..., x_N$ with real coefficients $w_1, ..., w_N$. It is no coincidence that this equation looks like a one-layer neural network (Jäkel et al., 2007), and because it is a linear combination of kernel functions these functions form a vector space.

There is a natural way to equip this vector space with an inner product. Let $g(x) = \sum_{i=1}^{M} v_i k(x, y_i)$ be another function from the vector space. An inner product between these functions can be defined as

$$\langle f, g \rangle = \sum_{i=1}^{N}\sum_{j=1}^{M} w_i v_j k(x_i, y_j). \tag{8}$$

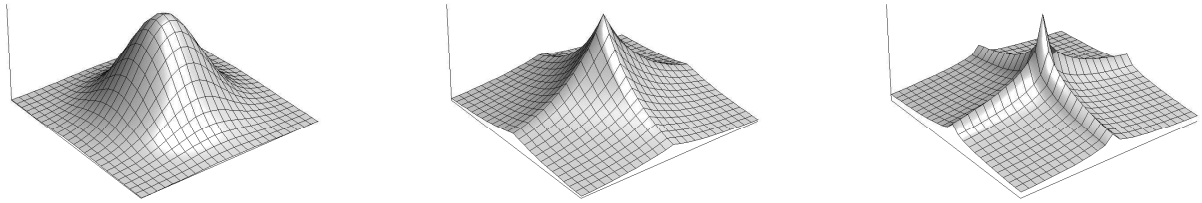

Figure 1. The similarity kernel for different values of p. From left to right: For $p = 2$ a Gaussian is obtained, for $p = 1$ a Laplacian is obtained, and for $p = \frac{1}{2}$ the axes are very prominent.

This can be shown to be well-defined and it is symmetric due to the symmetry of k. It is linear in its arguments, too, due to the linearity of the sum. To show that it is an inner product we need to make sure that it is also positive definite, that is $\langle f, f \rangle \geq 0$ and equality only holds for $f = 0$. Positivity is guaranteed by the defining property of a positive definite kernel k (4). Definiteness follows automatically for positive definite kernels but is a bit more difficult to see (Schölkopf & Smola, 2002).

The vector space with the inner product that we introduced is almost a Hilbert space. Hilbert spaces can be thought of as a generalization of Euclidean spaces with a dimension that may be infinite. In order to be a Hilbert space the space needs to be complete, and the space we constructed can be completed by including certain limit points (Schölkopf & Smola, 2002). This completed space is then called a *reproducing kernel Hilbert space* (RKHS). It is called "reproducing" because of the following property,

$$\langle k(\cdot, x), f \rangle = \sum_{i=1}^{N} w_i k(x, x_i) = f(x), \tag{9}$$

stating that the inner product between a function f and one of the similarity functions $k(\cdot, x)$ evaluates the function at x. Hence, when we take the inner product of two similarity functions

$$\langle k(\cdot, x), k(\cdot, y) \rangle = k(x, y), \tag{10}$$

the function $k(\cdot, y)$ is evaluated at x.

Remember that in Eq. (6) we decided to map each stimulus x to the vector space by applying the function $\Phi(x) = k(\cdot, x)$. Because of the reproducing property (10) the inner product of two stimuli x and y in the RKHS is given by their similarity:

$$\langle \Phi(x), \Phi(y) \rangle = \langle k(\cdot, x), k(\cdot, y) \rangle = k(x, y). \tag{11}$$

Calculating the similarity between two stimuli using a positive definite kernel k as given in (5) is therefore the same as taking the inner product in the Hilbert space that we constructed above. The similarity is given by an inner product as in the early work of Ekman (Ekman, 1954; Gregson, 1975; Borg & Groenen, 1997). Ironically, Shepard's suggestion to use the exponential as a link between distance and similarity has brought us back to the roots of MDS, the use of inner products.

The kernel metric

Like Euclidean space Hilbert space is a very rich structure with an inner product, a norm that is induced by the inner product and a metric that is induced by the norm. The norm of a function f in the Hilbert space is naturally defined as the square root of the inner product with itself $\|f\|^2 = \langle f, f \rangle$. In particular, for the similarity kernel (5) all stimuli are mapped to the unit sphere in Hilbert space:

$$\|\Phi(x)\|^2 = \langle \Phi(x), \Phi(x) \rangle = k(x, x) = \exp(0) = 1. \tag{12}$$

As all the points of the input space lie on the unit sphere in the Hilbert space the inner product is the cosine of the angle between the vectors in the Hilbert space.

Given a norm a natural definition of a metric is the norm of the difference vector. In the Hilbert space the distance between two functions f and g would then be given by the metric $\tilde{d}_p$ defined as $\tilde{d}_p = \|f - g\|$. Hence, the inner product in Hilbert space naturally induces a metric on the space via the norm:

$$\begin{aligned} \tilde{d}_p(x, y)^2 &= \|\Phi(x) - \Phi(y)\|^2 \\ &= \langle \Phi(x) - \Phi(y), \Phi(x) - \Phi(y) \rangle \\ &= \langle \Phi(x), \Phi(x) \rangle - 2 \langle \Phi(x), \Phi(y) \rangle + \langle \Phi(y), \Phi(y) \rangle \\ &= 2 - 2k(x, y) \\ &= 2 - 2 \exp(-d_p(x, y)^p) \end{aligned} \tag{13}$$

where we have used that the similarity kernel $k(x, x) = 1$ for all x. It is instructive to note that this new metric is a monotonic transform of d_p. In the Shepard-Kruskal multidimensional scaling procedure only ordinal properties of the data are used and therefore this new metric space is as good a representation for ordinal data as the ℓ_p space on which it is based. There are of course many more metric spaces that will do the same but do not have the structure of a Hilbert space. For some metric spaces there is no monotone metric transform such that they can be embedded into a Hilbert space, so it is not completely trivial to note that we can obtain a Hilbert space representation here (Lew, 1978).

Contrary to d_p the new metric is bounded from above. Points far apart in the space are separated by a distance which is at most $\sqrt{2}$. Psychologically this is an interesting property because it means that a stimulus that is already very different from another stimulus cannot become much more different. In fact, very often the notions of perceptual difference and similarity are only meaningful locally and measurements of large perceptual distances are not available. An example would be color space where it is easy to obtain local measurements of similarity, for instance by looking at discrimination thresholds. However, global measures are not easily available. If directly asked for a judgment of the dissimilarity of colors far apart in color space, typically subjects find themselves unable to express a more precise answer than "totally different" (Indow, 1994).

Triangle inequality

With regard to modeling perceptual similarity, it may seem that the kernel metric inherits all problems of the distance d_p on which it is based. Perhaps the problems may even seem worse because of the assumptions of the Hilbert space—but this is not so. Of course, the kernel metric fulfills the metric axioms and they have been subject to considerably criticism (Tversky, 1977). In some cases the empirical dissimilarity from a point to itself may not be zero and the symmetry of the dissimiliarities is not warranted. These violations may not always be explained by measurement noise and response biases. Symmetry, for example, can be violated if the comparison has a direction and one of the stimuli is more prototypical than the other, or receives more attention. Checking for violations of symmetry is relatively easy and even if an experimental measure is not completely symmetric, in practice it is often simply forced to be symmetric. Similarly, constant self-similarity is simply assumed in practice. In any case, whether one's data show symmetry and constant self-similarity, at least approximately, can easily be checked. The situation is not so simple for the triangle inequality. As dissimilarity measurements are usually only on an ordinal level, at most on an interval scale, for a finite set of points the triangle inequality can always be trivially satisfied by adding a big enough constant to the dissimilarity measurements. Hence, experimentally the triangle inequality can only be tested in conjunction with additional assumptions.

Tests of the triangle inequality

Assuming the ℓ_p norm, Shepard noted that concave (i.e., indented) iso-similarity contours lead to a violation of the triangle inequality (Shepard, 1964). Figure 2 shows the unit "balls" for the ℓ_p norm for different values of p assuming equal weights for both dimensions. All points on the curves have distance one to the center (in their respective norms). Figure 3 shows why the triangle inequality is not fulfilled for values of $p < 1$. For $p < 1$ the unit ball becomes concave. In this case, the distance from x to y is one, the distance from y to z is also one. Therefore, traveling from x to z via y takes two units but traveling directly, that is on a straight line, from x to z takes more than two units (the distance from x to w is greater than one and the distance from z to w is also greater than one). Psychologically, one would expect indented iso-similarity curves if subjects based their similarity judgments on matching dimensions. In his experiments Shepard found violations of the triangle inequality but he could attribute them to pooling subjects with different response strategies.

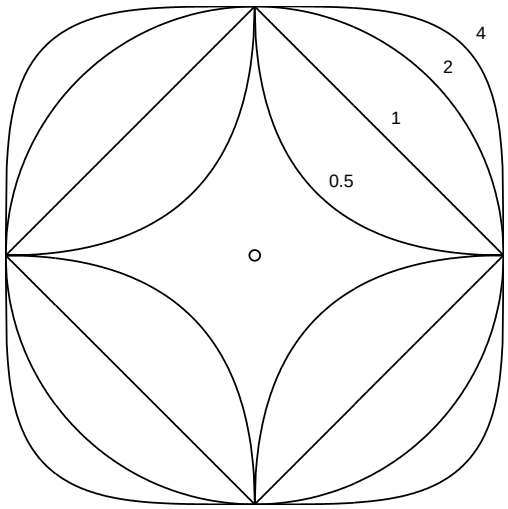

Figure 2. Unit balls of the ℓ_p norm for different values of p.

Tversky and Gati (1982) conducted a study that tested the triangle inequality (and matching behavior) more directly than Shepard (1964) did. Instead of assuming the ℓ_p norm they could devise testable predictions by only assuming the weaker assumption of segmental additivity (Beals et al., 1968; Blumenthal, 1953) in addition to the triangle inequality. By segmental additivity they meant the following: All pairs of points x and z can be joined by a segment (e.g., a straight line if the space is Euclidean) such that for any point w that is on the path between x an z the distance from x to z is exactly the sum of the distances from x to w and from w to z, $d(x, z) = d(x, w) + d(w, z)$. Implicitly we made this assumption above when we demonstrated that the distance d_p (2) does not fulfill the triangle inequality for $p < 1$. The assumption of segmental additivity is so intuitive that if it were to be given up the whole idea of representing similarity by geometric relations in a psychological space would seem to lose its intuitive appeal. Metrics with segmental additivity are a rather wide class of metrics. They include all Minkowski metrics (d_p with $p \geq 1$) and Riemannian curved geometries. Segmental additivity is one of the basic intuitions underlying MDS (Beals et al., 1968) and other, more recent embedding methods (Roweis & Saul, 2000; Tenenbaum, Silva, & Langford, 2000). Tversky and Gati found however that metrics with additive segments cannot account for their data.

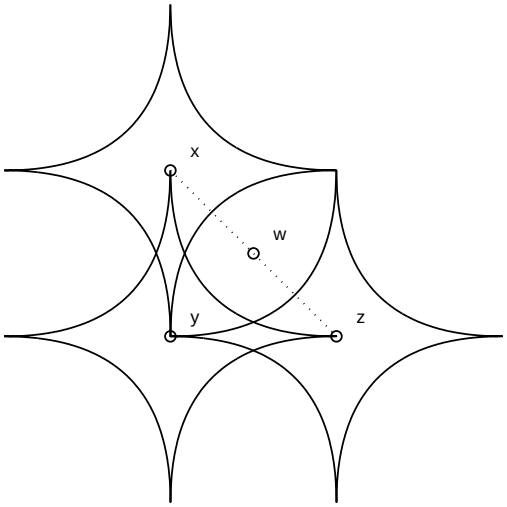

Figure 3. Violation of the triangle inequality for concave unit balls. The distances from x to y and from y to z are 1. Hence, traveling from x to z via y takes 2 units. Traveling from x to z directly via w takes more than 2 units as w is outside the unit balls of x and z.

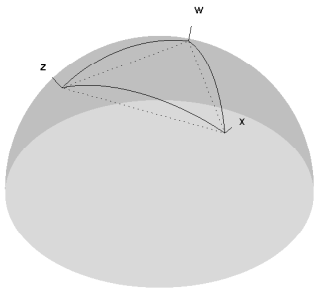

Figure 4. The similarity kernel maps the stimuli x, w and z from Figure 3 to the unit sphere in a Hilbert space. The distance from x to z is smaller than the sum of the distances from x to w and from w to z. The metric does not have additive segments in the original space because the distances are computed by the shortest connection in Hilbert space (the dotted lines).

Metrics without segmental additivity

These results have led Tversky and Gati, and many researchers after them, to prefer non-metric models of similarity. The contrast model (Tversky, 1977) is the most prominent example. Nevertheless, Tversky and Gati do acknowledge that the triangle inequality on its own is not constraining the class of similarity models very much. There are many metric models without segmental additivity that can be reconciled with their data, for example, the so-called "metric for bounded response scales" (Borg & Groenen, 1997). Another such metric was suggested by Tversky and Gati themselves and is given by the ℓ_p formula (2) taken to the power of p, which results in a metric for $0 < p \leq 1$, as already noted by Carroll and Wish (1974). Tversky and Gati (1982, p. 151) therefore conclude that the choice between non-metric models and metrics without segmental additivity is "more likely to be made on the basis of theoretical rather than empirical considerations". The kernel metric $\tilde{d}_p$ (13) that we introduced above is theoretically well-motivated by Shepard's law and also does not fulfill segmental additivity. Thus, the kernel metric may provide a theoretically well-founded metric alternative to the non-metric models that Tversky and Gati favor.

Note that the exponent in the definition of the kernel (5) is the p^{th} power of d_p. The iso-similarity curves of the kernel metric are identical to the iso-similarity curves of the ℓ_p norm (Figure 2) because the same value for the ℓ_p formula implies the same distance in the kernel metric. Contrary to the ℓ_p norm, the kernel metric can have indented iso-similarity curves and is therefore compatible with the data by Shepard (1964) and Tversky and Gati (1982). For a metric with indented unit balls it is exceedingly difficult to interpret a configuration of stimuli as depicted in Figure 3 as a map or any other intuitive geometry, "despite the natural tendency to do so" (Tversky & Gati, 1982, p. 151). However, as the kernel metric is a metric that is derived from an inner product we may use some of our Euclidean intuitions in its analysis.

The metric $\tilde{d}_p$ leads to segmental additivity in a higher dimensional space. The points in the psychological space are mapped to the unit sphere in the infinite dimensional Hilbert space (12). In Figure 4 we have depicted a three dimensional subspace that contains the points x, w and z from Figure 3 mapped to the Hilbert space using the mapping $\Phi(x) = k(\cdot, x)$. The locations of the three points in the figure are calculated from their inner products (that are given by the similarity kernel) such that the distance in RKHS is the same as the Euclidean distance in the three-dimensional space that is depicted. The metric $\tilde{d}_p$ is the metric of the Hilbert space in which the original psychological space is embedded and therefore the distance between two points is given by the chord that joins them (the dotted lines). Note that in the embedding space w is not on the way from x to z, it even lies in a different dimension in Figure 4. This is because any kernel matrix for Shepard's law will always have full rank, if no two points are identical. In fact, none of the points that lie on the chord that joins x and z can be a potential stimulus because we know that all stimuli from the original space lie on the unit sphere when mapped to the Hilbert space (12). Hence, segmental additivity does not hold in the original psychological space which is the one depicted in Figure 3 and also the one that Tversky and Gati examined. It is as if you can take a shortcut through the sphere in order

to get to another stimulus. You do not have to visit any other stimuli on the way. Any visit to another stimulus implies a detour. Hence, no two distinct points x and z in the original space can be joined by an additive segment in the original space because for all other points w in the original space it holds that $\tilde{d}_p(x,z) < \tilde{d}_p(x,w) + \tilde{d}_p(w,z)$.

Normally in MDS, if the stimulus space has a smaller dimension than the embedding space then the stimuli that are presented to a subject will fall onto a (non-linear) submanifold in the embedding space. The manifold has the dimension of the stimulus space. A simple example for this is the color circle (Shepard, 1980). If an experimenter chooses the one-dimensional set of stimuli that is comprised of only monochromatic lights then these stimuli will have to be embedded on a circle in two dimensions. The distance is given by the direct connection in the embedding space and not by the shortest path on the stimulus manifold (that only consists of monochromatic lights). In such a case it is no surprise that the metric does not fulfill segmental additivity when only the subset of monochromatic lights is considered. This chordal metric is in fact the standard example for a metric without additive segments (Beals et al., 1968). The kernel metric we have presented here is very similar to this example, it represents each stimulus on the unit sphere in a (high-dimensional) Hilbert space.[1]

Discussion

We have demonstrated how concerns about the triangle inequality that accompany all metric models of similarity can be addressed in a principled manner. It was important to remember that the experimental tests of the triangle inequality that Tversky and Gati (1982) reported were always in conjunction with a second assumption: segmental additivity. Shepard's law of generalization can be used to induce an inner product in a Hilbert space which in turn induces a metric with several psychologically appealing properties. It does not have additive segments and can be made consistent with the data by Tversky and Gati by choosing concave iso-similarity contours. It is also bounded from above and captures the intuition that similarity makes the most sense locally with only small changes in the stimulus. Stimuli far apart in perceptual space are merely completely different and more precise judgments of similarity are difficult.

There are other ways to address concerns about the triangle inequality for similarity. Within the framework of Minkowski metrics one could imagine that attention shifts and fluctuations not only occur over subjects and experimental contexts but also from trial to trial, possibly depending on the stimuli under consideration in each trial. Matching effects could thus be incorporated by giving matching dimensions a greater weight in the metric. Such attention shifts and fluctuations have frequently been suggested to account for non-metricity in metric models (Shepard, 1964; Micko & Fischer, 1970; Nosofsky, 1986; Tversky & Gati, 1982; Laub & Müller, 2004). Another suggestion has been to include spatial density as a factor in similarity judgments (Krumhansl, 1978). In contrast to all the metric models that one could entertain, the contrast model abandons the metric axioms altogether and directly tries to model matching behavior (Tversky, 1977).

Recently, Dzhafarov and Colonius (2007) have presented Dissimilarity Cumulation Theory as a principled way of constructing a metric from non-metric dissimilarity measurements. The major application of this theory has been to discrimination probabilities. However, also dissimilarity ratings of the kind considered by Tversky and Gati (1982) could potentially be used in their construction (Dzhafarov & Colonius, 2007, p.291). In addition, in an earlier work (Dzhafarov & Colonius, 2006) some of Shepard's generalization gradients have been analyzed and this led to a metric without additive segments, consistent with the metric suggested here. Here we have somehow naively assumed that similarity measurements as obtained by generalization gradients or by direct similarity ratings are monotonically related to each other. But the link may be more complicated and perhaps indented iso-similarity contours only occur for direct similarity judgments and not for generalization gradients (Nosofsky, 1986). Other connections between different operationalizations of similarity could be formalized within Dissimilarity Cumulation Theory. Dissimilarity Cumulation Theory is also sufficiently general to be able to deal with asymmetry and non-constant self-similarity, both of which we have neglected here.

In any case, the similarity component of many successful categorization models does assume Shepard's law (Nosofsky, 1986; Kruschke, 1992; Love et al., 2004, e.g.). There is a close correspondence between kernel methods, as they are used in machine learning and statistics, and exemplar models of categorization (Ashby & Alfonso-Reese, 1995; Jäkel, Schölkopf, & Wichmann, 2008). It is, in fact, this correspondence that motivated this work. One should expect that matching effects like (Tversky & Gati, 1982) observed play a role in categorization, too. (Nosofsky, 1986) did not find it necessary to incorporate matching effects for his highly confusable stimuli. Nevertheless, it has been suggested that matching behavior should be incorporated into models of categorization (Verguts, Ameel, & Storms, 2004). This can be done in exemplar models by setting $p < 1$. If p is simply seen as a free parameter of the exemplar model it will be hardly surprising that this can be done. However, it is perhaps interesting to note that it can be done without giving up the metric axioms.

References

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology, 39*, 216-233.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37*, 372-400.

[1] One can also analyze the intrinsic geometry of the manifolds that are given by a kernel. Perhaps surprisingly, even though Shepard's law demands that all points lie on the unit sphere the geometry can be flat (Burges, 1998).

Attneave, F. (1950). Dimensions of similarity. *American Journal of Psychology*, *63*, 516-556.

Beals, R., Krantz, D. H., & Tversky, A. (1968). Foundations of multidimensional scaling. *Psychological Review*, *75*, 127-142.

Blumenthal, L. M. (1953). *Theory and applications of distance geometry*. Oxford: Clarendon Press.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications*. New York: Springer.

Burges, C. (1998). Geometry and invariance in kernel based methods. In B. Schölkopf, C. Burges, & A. Smola (Eds.), *Advances in kernel methods: Support vector learning* (p. 89-116). Cambridge, MA: MIT Press.

Carroll, J. D., & Wish, M. (1974). Multidimensional perceptual models and measurement methods. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. 2). New York: Academic Press.

Chater, N., & Vitanyi, P. M. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology*, *47*, 346-369.

Dzhafarov, E. N., & Colonius, H. (2006). Reconstructing distances among objects from their discriminability. In H. Colonius & E. N. Dzhafarov (Eds.), *Measurement and representation of sensations*. Mahwah, NJ: Erlbaum Publ. Co.

Dzhafarov, E. N., & Colonius, H. (2007). Dissimilarity cumulation theory and subjective metrics. *Journal of Mathematical Psychology*, *51*, 290-304.

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences*, *21*, 449-498.

Ekman, G. (1954). Dimensions of color vision. *Journal of Psychology*, *38*, 467-474.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

Gregson, R. A. M. (1975). *Psychometrics of similarity*. New York: Academic Press.

Indow, T. (1994). Metrics in color spaces: Im Kleinen und im Großen. In G. H. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology.* New York: Springer.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2007). A tutorial on kernel methods for categorization. *Journal of Mathematical Psychology*, *51*, 343-358.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008). Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin & Review*, *15*(2), 256-271.

Koldobsky, A., & Koenig, H. (2001). Aspects of isometric theory of Banach spaces. In W. B. Johnson & J. Lindenstrauss (Eds.), *Handbook of the geometry of Banach spaces* (p. 899-939). Amsterdam: Elsevier.

Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, *85*(5), 445-463.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22-44.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1-27.

Laub, J., & Müller, K.-R. (2004). Feature discovery in non-metric pairwise data. *Journal of Machine Learning Research*, *5*, 801-818.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*, 1-15.

Lew, J. S. (1978). Some counterexamples in multidimensional scaling. *Journal of Mathematical Psychology*, *17*, 247-254.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological Review*, *111*(2), 309-32.

Micko, H. C., & Fischer, W. (1970). The metric of multidimensional psychological spaces as a function of the differential attention to subjective attributes. *Journal of Mathematical Psychology*, *7*, 118-143.

Nosofsky, R. M. (1986). Attention, similarity, and the indentification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39-57.

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology*, *34*, 393-418.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*, 2323-2326.

Schoenberg, I. J. (1938). Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, *44*(3), 522-536.

Schölkopf, B., & Smola, A. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325-345.

Shepard, R. N. (1958). Stimulus response generalization: Deduction of the generalization gradient from a trace model. *Psychological Review*, *65*(4), 242-256.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika*, *27*(2), 125-140.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, *1*, 54-87.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390-398.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317-1323.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629-640.

Tenenbaum, J. B., Silva, V. de, & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*, 2319-2323.

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika*, *17*, 401-419.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327-352.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, *89*(2), 123-154.

Tversky, A., & Krantz, D. H. (1970). The dimensional representation and the metric structure of similarity data. *Journal of Mathematical Psychology*, *7*, 572-596.

Verguts, T., Ameel, E., & Storms, G. (2004). Measures of similarity in models of categorization. *Memory & Cognition*, *32*(3), 379-89.

Young, G., & Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, *3*, 19-21.