Available online at www.sciencedirect.com
ScienceDirect
Mathematical Social Sciences 55 (2008) 107–115
mathematical social sciences
www.elsevier.com/locate/econbase

Axiomatization of an exponential similarity function ☆

Antoine Billot [a,b,c], Itzhak Gilboa [d,e,*], David Schmeidler [d,f]

[a] University de Paris II, France
[b] PSE-Jourdan, France
[c] CERAS–ENPC, France
[d] Tel-Aviv University, Israel
[e] HEC-Paris, France
[f] The Ohio State University, United States

Received 2 May 2007; accepted 2 August 2007
Available online 14 August 2007

Abstract

An individual is asked to assess a real-valued variable y based on certain characteristics $x=(x^1,\ldots, x^m)$, and on a database consisting of n observations of $(x^1,\ldots, x^m, y)$. A possible approach to combine past observations of x and y with the current values of x to generate an assessment of y is *similarity-weighted averaging*. It suggests that the predicted value of y, y^s_{n+1}, be the weighted average of all previously observed values y_i, where the weight of y_i is the similarity between the vector $x^1_{n+1},\ldots, x^m_{n+1}$, associated with y_{n+1}, and the previously observed vector, $x^1_i,\ldots, x^m_i$. This paper axiomatizes, in terms of the prediction y_{n+1}, a similarity function that is a (decreasing) exponential in a norm of the difference between the two vectors compared.

Keywords: Similarity function; Axiom; Exponential decay

1. Introduction

In many prediction and learning problems, an individual attempts to assess the value of a real variable y based on the values of relevant variables, $x=(x^1,\ldots, x^m)$, and on a database, B,

☆We wish to thank Jerome Busemeyer for comments and references. Gilboa and Schmeidler gratefully acknowledge support from the Polarization and Conflict Project CIT-2-CT-2004-506084 funded by the European Commission-DG Research Sixth Framework Programme and from the Israel Science Foundation (Grant Nos. 790/00 and 975/03).
*Corresponding author. Tel-Aviv University, Israel.
E-mail addresses: billot@u-paris2.fr (A. Billot), igilboa@post.tau.ac.il (I. Gilboa), schmeid@tau.ac.il (D. Schmeidler).

0165-4896/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.mathsocsci.2007.08.002

consisting of past observations of the variables $(x_i, y_i)=(x_i^1,\ldots, x_i^m, y_i)$, $i=1,\ldots, n$. Some examples for the variable y include the weather, the behavior of other people, and the price of an asset. The relevant variables x may represent meteorological conditions, psychosocial cues, or the attributes of the asset, respectively.

There are many well-known approaches for the prediction of y given x and the database B. For instance, regression analysis is such a method. k-nearest neighbor techniques (Fix and Hodges, 1951, 1952) would be another method, predicting the value of y at a point x by the values that y has assumed for points close to x. In fact, the literature in statistics and in machine learning offers a variety of methods for this problem, which encompasses a wide spectrum of problems that people encounter in their daily lives as well as in professional endeavors.

One approach to deal with the classical learning/prediction problem is to use a similarity-weighted average: fix a similarity function $s : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_{++}$ and, given the database B and the new data point $x\in\mathbb{R}^m$, generate the prediction

$$y^s = \frac{\sum_{i\le n} s(x_i, x)y_i}{\sum_{i\le n} s(x_i, x)}.$$

This formula was suggested and axiomatized in Gilboa, Lieberman, and Schmeidler, 2006 .[1,2] They assume that, for every $n\in\mathbb{N}_{++}$, any database B (consisting of $n\ge 1$ observations in $\mathbb{R}^{m+1}$), and every new point $x\in\mathbb{R}^m$, a predictor has an ordering over $\mathbb{R}$, $\succsim_{B,x}$, interpreted as "more likely than". They show that these orderings satisfy certain axioms if and only if there exists a similarity function such that the ordering ranks possible predictions y according to their proximity to y^s.

In this paper, we investigate the explicit form of the similarity function s, in the context of the similarity-weighted formula. That is, we assume that Y is assessed according to

$$Y(B, x) = \frac{\sum_{i\le n} s(x_i, x)y_i}{\sum_{i\le n} s(x_i, x)} \tag{1}$$

where the function $Y(\cdot,\cdot)$ is defined on the all databases, $\mathbb{B} = \cup_{n\ge 1}(\mathbb{R}^{m+1})^n$, and for all $x\in\mathbb{R}^m$. The derivation of formula (1) by Gilboa et al. (2006) is done for each x separately, considering the rankings of possible values of $Y(B, x)$ for various databases B, but for a fixed $x\in\mathbb{R}^m$. Hence, they obtain a separate function $s(\cdot, x)$ for each x. This function is strictly positive and it is unique up to multiplication by a positive number. For concreteness, we here normalize this function such that $s(x, x)=1$ for every x. With this convention, s is unique.

We consider the behavior of $Y(\cdot,\cdot)$ when one varies its arguments. We suggest certain consistency conditions on Y, referred to as "axioms", which characterize an exponential functional form, namely, a similarity function s that satisfies, for every $x, z\in\mathbb{R}^m$,

$$s(z, x) = \exp[-\boldsymbol{\nu}(x - z)] \tag{2}$$

for some norm $\boldsymbol{\nu}$ on $\mathbb{R}^m$. Assuming that the assessments Y are observable, our result may be interpreted as showing what observable implications are there to the assumption of exponential similarity (2) in the context of the similarity-weighted average formula (1).

[1] It is reminiscent of derivations in Gilboa and Schmeidler (2003) and in Billot et al. (2005). It also bears resemblance to kernel-based methods of estimations, as in Akaike (1954), Rosenblatt (1956), Parzen (1962) and others. See Silverman (1986) and Scott (1992) for surveys.

[2] The term similarity at this point does not impose any restriction on the function. It just indicates that this function is used in a formula like the one above.

The notion of a similarity function which is decaying exponentially as a function of distanceis rather natural, and appears in other contexts as well. For instance, Shepard (1987) derives an exponential similarity function which measures the probability of generalizing a response from one stimulus to another. An exponential decay function is used to model the probability of recall (see, for instance, Bolhuis, Bijlsma, and Ansmink, 1986), which may be interpreted as a measure of the similarity between two points of time. The present paper shows that exponential decay, relative to some norm, has, and is characterized by rather appealing properties also when similarity is used for the computation of similarity-weighted average as in Eq. (1).

The axioms and the main result are stated in the next section. They are followed by comments on several special cases of the norm $\boldsymbol{\nu}$, the special case of a single-dimensional space, and a general discussion. Proofs are to be found in an Appendix.

2. Main result

Suppose that there are given functions $Y : \mathbb{B} \times \mathbb{R}^m \to \mathbb{R}$ and $s : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_{++}$ as in formula (1) (The positive integer m is fixed throughout the paper.). We impose the following axioms on Y:

A1. Shift Invariance

For every $B = (x_i, y_i)_{i \le n} \in \mathbb{B}$, and every $x, w \in \mathbb{R}^m$,

$$Y((x_i + w, y_i)_{i \le n}, x + w) = Y((x_i, y_i)_{i \le n}, x).$$

A1 states that the prediction does not depend on the absolute location of the points (x_i), x in $\mathbb{R}^m$, but only on their relative location. More precisely, it demands that a shift in all independent variables in the database, accompanied by the same shift in the new independent variable for which prediction is required, will not affect the predicted value Y.

The next axiom requires that evidence that was obtained for further points has lower impact. It is restricted to a rather uncontroversial definition of "being further away": it is only required to hold along rays emanating from zero, when prediction is required for the point $x = \mathbf{0}$ (To avoid confusion we will denote the origin in $\mathbb{R}^m$ by a bold 0, $\mathbf{0}$.).

A2. Ray Monotonicity

For every $x, z \in \mathbb{R}^m$, $Y(((\lambda x, 1), (z, -1)), \mathbf{0})$ is strictly decreasing in $\lambda \ge 0$.

A2 considers databases consisting of two points, one, λx, at which the value 1 was observed, and another, z, at which the value -1 was observed. Obviously, Eq. (1) would generate a value Y in $(-1, 1)$ for such a database. When we vary λ, the value of Y will be higher, the more similar is λx considered to be to $\mathbf{0}$. A2 states that, if we move λx further away from $\mathbf{0}$ (along the ray through x), it will be considered less similar to $\mathbf{0}$, and the prediction Y will decrease (i.e., it will move away from 1 toward -1.).

A3. Symmetry

For every $x \in \mathbb{R}^m$,

$$Y(((x, 1), (\mathbf{0}, 0)), \mathbf{0}) = Y(((\mathbf{0}, 1), (x, 0)), x).$$

A3 considers two situations. In the first, one has observed the value 1 for $x \in \mathbb{R}^m$, and the value 0 for $\mathbf{0} \in \mathbb{R}^m$, and one is asked to make a prediction for $\mathbf{0} \in \mathbb{R}^m$. In the second situation, the roles

are reversed: the value 1 was observed at $\mathbf{0}\in\mathbb{R}^m$, the value 0 at $x\in\mathbb{R}^m$, and the prediction is requested for x. A3 then requires that the prediction be the same in these two situations. Intuitively, it demands that the impact an observation at x has on observation at $\mathbf{0}$ is the same as the impact of the same observation at $\mathbf{0}$ has on observation at x.

Axiom 4 is reminiscent of A1, but the antecedent is more restrictive and the conclusion stronger. It applies to a database where all the independent variables are on a ray through the origin. A shift along this ray leaves the prediction unchanged although the independent variable for which the predictions are made is the origin before and after the shift. Formally,

A4. Ray Shift Invariance

Let there be given $B=(\alpha_i\upsilon,y_i)_{i\le n}\in\mathbb{B}$, for some $\upsilon\in\mathbb{R}^m$ and $\alpha_i\ge 0$ $(i\le n)$. Then, for every $\theta>0$, $Y((\alpha_i\upsilon+\theta\upsilon,\, y_i)_{i\le n},\, \mathbf{0})=Y((\alpha_i\upsilon,\, y_i)_{i\le n},\, \mathbf{0})$.

Our last axiom is,

A5. Self-relevance

For every $x,z\in\mathbb{R}^m$,

$$Y(((\mathbf{0},1),(x,0)),z)\le Y(((\mathbf{0},1),(x,0)),\mathbf{0}).$$

A5 considers a simple database B consisting of two points: the value 1 was observed for the point $\mathbf{0}$, while the value 0 was observed for the point x. Given such a database, any prediction generated by Eq. (1) is necessarily in [0,1]. Intuitively, the prediction generated given this database, for every z, is higher the higher is the similarity of z to 0 relative to its similarity to x. Self-relevance requires that this relative similarity be maximized at $z=0$. That is, no other point $z\neq 0$ can be more similar to 0 than to x, as compared to 0 itself.

Recall that a *norm* on $\mathbb{R}^m$ is a function $\boldsymbol{\nu}:\mathbb{R}^m\to\mathbb{R}_+$ satisfying:

(i) $\boldsymbol{\nu}(\xi)=0$ iff $\xi=0$;
(ii) $\boldsymbol{\nu}(\lambda\xi)=|\lambda|\boldsymbol{\nu}(\xi)$ for all $\xi\in\mathbb{R}^m$ and $\lambda\in\mathbb{R}$;
(iii) $\boldsymbol{\nu}(\xi+\zeta)\le\boldsymbol{\nu}(\xi)+\boldsymbol{\nu}(\zeta)$ for all $\xi,\zeta\in\mathbb{R}^m$.

We can now state our main result:

Theorem 1. *Let there be given a function Y as in formula (1), were s is normalized by: s(x, x) = 1 for all $x\in\mathbb{R}^m$. The following are equivalent:*

(i) Y satisfies A1–A5;
(ii) There exists a norm $\boldsymbol{\nu}:\mathbb{R}^m\to\mathbb{R}_+$ such that
(∗) $s(x,z)=\exp[-\boldsymbol{\nu}(x-z)]$ for every $x,z\in\mathbb{R}^m$

We observe that, given s, the norm $\boldsymbol{\nu}$ is uniquely defined by (∗), and vice versa.

The shift axioms (A1) enables us to state the rest of the axioms for $Y(\cdot,0)$ rather than for $Y(\cdot,w)$ for every $w\in\mathbb{R}^m$. As will be clear from the proof of the theorem, one may drop A1, strengthen the other axioms so that they hold for every $w\in\mathbb{R}^m$, and obtain a similar representation that depends on a more general distance function (that is not necessarily based on a norm).

It will also be clear from the proof that our result can be generalized at no cost to the case that the data points x_i belong to any linear space (rather than $\mathbb{R}^m$). This is true also of the axiomatization in Gilboa et al. (2006). Taken together, the two results may be viewed as axiomatically deriving a norm on a linear space, based on predictions Y.

The similarity function obtained in Gilboa et al. (2006) has no structure whatsoever. The only property that follows from their axiomatization is the positivity of s. An important feature of our result is that observable conditions on predictions Y imply that $\boldsymbol{\nu}$ is a norm, and this, in turn, imposes restrictions on the similarity function. For example, since for a norm $\boldsymbol{\nu}$, $\boldsymbol{\nu}(\xi)=\boldsymbol{\nu}(-\xi)$, we conclude that $s(x, z)=s(z, x)$, that is, that s is symmetric.

Another important feature of norms is that they satisfy the triangle inequality. This would imply that s satisfies a certain notion of transitivity. Specifically, it is not hard to see that, given the representation (∗), the triangle inequality for $\boldsymbol{\nu}$ implies thats satisfies multiplicative transitivity, namely, for every $x, z, w \in \mathbb{R}^m$,

$$s(x,w) \geq s(x,z)s(z,w).$$

Thus, if both x and w are similar to z to some degree, x and w have to be similar to each other to a certain degree. Specifically, if both $s(x, z)$ and $s(w, z)$ are at least ε, then $s(x, w)$ is bounded below by ε^2.

3. Special cases

One may impose additional conditions on Y that would restrict the norm that one obtains in the theorem. For instance, consider the following axiom:

A6. Rotation

Let P be an $m \times m$ orthonormal matrix. Then, for every $B=(x_i, y_i)_{i\leq n}$, $Y((x_i, y_i)_{i\leq n}, \mathbf{0})= Y((x_iP, y_i)_{i\leq n}, \mathbf{0})$.

A6 asserts that rotating the database around the origin would not change the prediction at the origin. It is easy to see that in this case the norm $\boldsymbol{\nu}$ coincides with the standard norm on $\mathbb{R}^m$.

For certain applications, one may prefer a norm that is defined by a weighted Euclidean distance, rather than by the standard one. To obtain a derivation of such a norm, we need an additional definition.

For two points $z, z' \in \mathbb{R}^m$, we write $x \sim x'$ if the following holds: for every $B \in \mathbb{B}$, and $y \in \mathbb{R}$, $Y((B, (x, y)), \mathbf{0}) = Y((B, (x', y)), \mathbf{0})$, where $(B, (x, y))$ denotes the database obtained by concatenation of B with (x, y). In light of Eq. (1), it is easy to see that two vectors x and x' are considered $\sim$-equivalent if and only if $s(x, \mathbf{0})=s(x', \mathbf{0})$. Using this fact, or using the definition directly, one may verify that $\sim$ is indeed an equivalence relation.

In the presence of axiom A1, two vectors x and x' are considered $\sim$-equivalent if observing y at a point that is x-removed from the new point has the same impact on the prediction as observing y at a point that is x'-removed from the new point.

For $j \leq m$, let $e_j \in \mathbb{R}^m$ be the j-th unit vector in $\mathbb{R}^m$ (that is, $e_j^k=1$ for $k=j$ and $e_j^k=0$ for $k\neq j$). we can now state.

A7. Elliptic Rotation

Assume that, for j, $k \leq m$ and $\beta>0$, $e_j \sim \beta e_k$. Let θ, $\mu>0$ be such that $\beta\theta^2+\mu^2=\beta$. Then for every $x=(x^1,\ldots, x^m)$, $x+e_j \sim x+\theta e_j+\mu e_k$.

A7 requires that $\sim$-equivalence classes would be elliptic. Specifically, it compares a unit vector on the j-th axis to a multiple of the unit vector on the k-axis. It assumes that β is the appropriate multiple of e_k that would make it equivalent to e_j. It then considers the ellipse connecting these points, and demands that this ellipse would lie on an equivalence curve of $\sim$. It can be verified that A7 will imply that $\boldsymbol{\nu}$ is defined by a weighted Euclidean distance.

More generally, one may use the equivalence relation above to state axioms that correspond to various specific norms. In particular, any L_p norm can be derived from an axiom that parallels A7.

4. A single dimension

An interesting special case is where there is only one predictor, i.e., when $m=1$. A prominent example would be when the data are indexed by time. In this case, the point for which a prediction is required is larger, that is, further into the future, than any point in the database and not all the axioms are needed for our main result. Moreover, when m = 1 the exponential similarity function can also be justified on different grounds. We begin by stating the appropriate versions of the axioms.

Let $\mathbb{B}' = \{((x_i,y_i)_{i\le n})|(x_i,y_i)\in \mathbb{R}^2, x_i \ge x_j \text{ for } i>j\}$. Denote by $\mathbb{B}'_0$ the union of $\mathbb{B}'$ and the set containing the empty database (corresponding to $n=0$). Assume that Y is defined on

$$\mathbb{D} = \{((x_i,y_i)_{i\le n},x)|(x_i,y_i)_{i\le n}\in\mathbb{B}', x\in\mathbb{R}, x\ge x_n\}.$$

Re-write the axioms as follows.

A1'. Shift Invariance

For every $((x_i,y_i)_{i\le n},x)\in\mathbb{D}$, and every $w\in\mathbb{R}$, $Y((x_i+w, y_i)_{i\le n}, x+w) = Y((x_i, y_i)_{i\le n}, x)$.

A2'. Monotonicity

$Y(((-1, 1), (\lambda, -1)), 1)$ is strictly decreasing in $\lambda\in[-1, 1]$.

A4'. Ray Shift Invariance

For every $((x_i,y_i)_{i\le n},x)\in\mathbb{D}$, and every $w\ge 0$, $Y((x_i, y_i)_{i\le n}, x+w) = Y((x_i, y_i)_{i\le n}, x)$.

The Shift Invariance axiom states that shifting the entire database, as well as the new point, does not affect the prediction. The monotonicity axiom states that the closer is a datapoint (λ) to the new prediction (1), the higher is its impact, that is, the -1 associated with λ has a greater weight in the prediction for $x=1$ as compared to another datapoint (1 observed at -1). Finally, the Ray Shift Invariance states that if a prediction is required for a later point ($x+w$ rather than x), but no new datapoint have been observed, the prediction does not change.

Interpreting the single predictor as time, the axioms have quite intuitive justifications: Shift Invariance states that the point at which we start measuring time is immaterial. Monotonicity simply requires that a more recent experience have a greater impact on current predictions. Finally, Ray Shift Invariance can be viewed as stating that the predictor does not change her prediction simply because time has passed. If no new datapoints were added, no change in prediction would result.

In a single dimension, the exponential similarity function allows one to summarize a database by a single case, such that, for all future observations and all future prediction problems, the

summary case would serve just as well as the entire database. Specifically, we formulate a new condition:

Summary.

For every $((x_i, y_i)_{i\leq n})\in\mathbb{B}$, there exists $(\overline{x},\overline{y})\in\mathbb{R}^2$, such that for every $((x'_i, y'_i)_{i\leq m})\in\mathbb{B}_0$ with $x'_1\geq x_n$ (if $m>0$), and every $x\geq x'_m$, $Y(((x_i, y_i)_{i\leq n}, (x'_i, y'_i)_{i\leq m}), x)=Y(((\overline{x},\overline{y}), (x'_i, y'_i)_{i\leq m}), x)$.

We can now state:

Proposition 2.

Let there be given a function Y as in formula (1), were s is normalized by: s(x, x) = 1 for all $x\in\mathbb{R}^m$. The following are equivalent:

(i) Y satisfies A1', A2', A4';
(ii) Y satisfies A1', A2', and Summary;
(iii) There exists $\theta\in\mathbb{R}_+$ such that $s(x, z)=\exp[-\boldsymbol{\theta}(z-x)]$ for every $z\geq x$.

Appendix A. Proof

Proof of Theorem 1. It is convenient to prove that (i) is equivalent to (ii) by imposing one axiom at a time. This will also clarify the implication of A1, A1 and A2, etc.[3]

It is easy to see that A1 is equivalent to the existence of a function $f:\mathbb{R}^m\to\mathbb{R}_{++}$, with $f(\mathbf{0})=1$, such that $s(x, z)=f(x-z)$ for every $x, z\in\mathbb{R}^m$. Indeed, if such an f exists, A1 will hold. Conversely, if A1 holds, one may define $f(x)=s(x, \mathbf{0})$ and use the shift axiom to verify that $s(x, z)=f(x-z)$ holds for every $x, z\in\mathbb{R}^m$.

Next consider A2. Since $f(x)=s(x, \mathbf{0})$, it is easy to see that A2 holds if and only if f is strictly decreasing along any ray emanating from the origin. Explicitly, A1 and A2 hold if and only if $s(x, z)=f(x-z)$ for every $x, z\in\mathbb{R}^m$ and $f(\lambda x)$ is strictly decreasing in $\lambda\geq 0$ for every $x\in\mathbb{R}^m, x\neq 0$ and $f(\mathbf{0})=1$.

It is easily seen that symmetry (A3) is equivalent to the fact that $f(x)=f(-x)$ for every $x\in\mathbb{R}^m$.

We now turn to A4. Consider a ray originating from the origin, $\{\lambda x|\lambda\geq 0\}$, for a given $x\in\mathbb{R}^m$ ($x\neq\mathbf{0}$). We observe that for Ray Invariance to hold, in the presence of Monotonicity, $s(\lambda x, \mathbf{0})$ has to be exponential in λ. To see this, observe that Ray Invariance implies that the ratio $s(k\lambda x, \mathbf{0})/s((k+1)\lambda x, \mathbf{0})$ is independent of k for every λ. This guarantees that $s(\lambda x, \mathbf{0})$ is exponential on the rational values of λ. Given monotonicity (A2) we conclude that for every $x\in\mathbb{R}^m$ there exists a number $\boldsymbol{\nu}_x$ such that $s(\lambda x, \mathbf{0})=\exp[-\lambda\boldsymbol{\nu}_x]$. Obviously, $\boldsymbol{\nu}\lambda_x=\lambda\boldsymbol{\nu}_x$ for $\lambda\geq 0$. A2 also implies that $\boldsymbol{\nu}_x>0$ for $x\neq\mathbf{0}$.

Combining these observations with the previous ones, we conclude that A1–A4 are equivalent to the existence of a function $f:\mathbb{R}^m\to\mathbb{R}_{++}$, such that $s(x, z)=f(x-z)$ for every $x, z\in\mathbb{R}^m$, where $f(\mathbf{0})=1$, $f(x)=f(-x)$ for every $x\in\mathbb{R}^m$, and, for every $x\in\mathbb{R}^m$ there exists a non-negative number $\boldsymbol{\nu}_x$ such that $f(x)=\exp[-\lambda\boldsymbol{\nu}_x]$ and $\boldsymbol{\nu}\lambda_x=\lambda\boldsymbol{\nu}_x$ for $\lambda\geq 0$. Further, $\boldsymbol{\nu}_x=0$ only for $x=\mathbf{0}$. Defining $\boldsymbol{\nu}(x)=\boldsymbol{\nu}_x$ we obtain the representation (∗) for a function $\boldsymbol{\nu}$ that satisfies all the conditions of a norm, apart from the triangle inequality.

[3] We will follow the order A1–A4. The exact implication of each subset of axioms separately can be similarly analyzed.

To conclude the proof, we need to show that $\boldsymbol{\nu}$ satisfies $\boldsymbol{\nu}(x+z) \leq \boldsymbol{\nu}(x)+\boldsymbol{\nu}(z)$ if and only if A5 holds. Consider arbitrary $x, z \in \mathbb{R}^m$. A5 states that

$$Y(((\mathbf{0},1),(x,0)),z) \leq Y(((\mathbf{0},1),(x,0)),\mathbf{0})$$

which implies that

$$\frac{s(\mathbf{0},z)}{s(\mathbf{0},z)+s(x,z)} \leq \frac{s(\mathbf{0},\mathbf{0})}{s(\mathbf{0},\mathbf{0})+s(x,\mathbf{0})}$$

or

$$\frac{s(\mathbf{0},z)}{s(\mathbf{0},z)+s(x,z)} \leq \frac{1}{1+s(x,\mathbf{0})}.$$

Equivalently, we have

$$\frac{s(\mathbf{0},z)+s(x,z)}{s(\mathbf{0},z)} \geq 1+s(x,\mathbf{0})$$

which is equivalent, in turn to

$$\frac{s(x,z)}{s(\mathbf{0},z)} \geq s(x,\mathbf{0})$$

and to

$$s(x,z) \geq s(x,\mathbf{0})s(\mathbf{0},z).$$

Observe that A5 is equivalent to this form of multiplicative transitivity independently of the other axioms. While we obtain the multiplicative transitivity condition only at $\mathbf{0}$, an obvious strengthening of A5 will imply that $s(x, z) \geq s(x, w)s(w, z)$ for every $x, z, w \in \mathbb{R}^m$.

Using the representation of s, we conclude that A5 is equivalent to the claim that, for every $x, z \in \mathbb{R}^m$,

$$\exp[-\boldsymbol{\nu}(x-z)] \geq \exp[-\boldsymbol{\nu}(x)-\boldsymbol{\nu}(-z)]$$

or

$$\boldsymbol{\nu}(x-z) \leq \boldsymbol{\nu}(x)+\boldsymbol{\nu}(-z).$$

Setting $\xi = x$ and $\zeta = -z$, we conclude that A5 holds if and only if $\boldsymbol{\nu}$ satisfies the triangle inequality.

This completes the proof of the theorem. □

Proof of Proposition 2. The equivalence of (i) and (iii) is proved as in the general case (see the Proof of Theorem 1 above). We wish to show that Summary may replace A4'. First, observe that Summary is a stronger condition than is A4'. This follows from restricting Summary to the case $m=0$, and observing that $Y((\bar{x}, \bar{y}), x) = \bar{y}$ for all x. Conversely, it is easy to verify that (iii) implies Summary. □

References

Akaike, H., 1954. An approximation to the density function. Annals of the Institute of Statistical Mathematics 6, 127–132.
Billot, A., Gilboa, I., Samet, D., Schmeidler, D., 2005. Probabilities as similarity-weighted frequencies. Econometrica 73, 1125–1136.
Bolhuis, J.J., Bijlsma, S., Ansmink, P., 1986. Exponential decay of spatial memory of rats in a radial maze. Behavioral and Neural Biology 46, 115–122.
Fix, E., Hodges, J., 1951. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. Technical Report 4, Project Number 21-49-004. USAF School of Aviation Medicine, Randolph Field, TX.
Fix, E., Hodges, J., 1952. Discriminatory Analysis: Small Sample Performance. Technical Report 21-49-004. USAF School of Aviation Medicine, Randolph Field, TX.
Gilboa, I., Schmeidler, D., 2003. Inductive inference: An axiomatic approach. Econometrica 71, 1–26.
Gilboa, I., Lieberman, O., Schmeidler, D., 2006. Empirical Similarity. Review of Economics and Statistics, 88, 433–444.
Parzen, E., 1962. On the estimation of a probability density function and the mode. Annals of Mathematical Statistics 33, 1065–1076.
Rosenblatt, M., 1956. Remarks on some nonparametric estimates of a density function. Annals of Mathematical Statistics 27, 832–837.
Scott, D.W., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley and Sons, New York.
Shepard, R.N., 1987. Toward a universal law of generalization for psychological science. Science 237, 1317–1323.
Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.