Mathematical Social Sciences 66 (2013) 346–351

Contents lists available at ScienceDirect

Mathematical Social Sciences

journal homepage: www.elsevier.com/locate/econbase

Asymmetric empirical similarity

Joshua C. Teitelbaum [*]

Georgetown University Law Center, 600 New Jersey Avenue NW, Washington, DC 20001, United States

ARTICLE INFO

Article history:
Received 21 January 2013
Received in revised form
19 July 2013
Accepted 26 July 2013
Available online 6 August 2013

ABSTRACT

The paper suggests a similarity function for applications of empirical similarity theory in which the notion of similarity is asymmetric. I propose defining similarity in terms of a quasimetric. I suggest a particular quasimetric and explore the properties of the empirical similarity model given this function. The proposed function belongs to the class of quasimetrics induced by skewed norms. Finally, I provide a skewness axiom that, when imposed in lieu of the symmetry axiom in the main result of Billot et al. (2008), characterizes an exponential similarity function based on a skewed norm.

1. Introduction

Similarity-based reasoning is an important topic in a wide range of fields, including artificial intelligence, cognitive science, decision theory, economics, and jurisprudence. Also known as analogical reasoning or case-based reasoning, similarity-based reasoning entails reasoning by analogy to past cases. A similarity-based reasoner evaluates the similarity between past cases and the case at hand and reaches a decision through application of the principle that like cases should be treated alike.

Gilboa et al. (2006) and Billot et al. (2005) suggested an axiomatic theory of similarity-based reasoning for real-valued assessment problems, known as empirical similarity theory.[1] Under this theory, assessments are made according to similarity-weighted averages of prior assessments. In particular, the theory posits that, given a new data point $x \in \mathbb{R}^n$ and a database of prior cases $(x_i, y_i)_{i \le t}$, $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ for all $i \le t$, one assesses the value of a real variable y according to the formula $y = \sum_{i \le t} s(x_i, x)y_i / \sum_{i \le t} s(x_i, x)$, where $s : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{++}$, known as the similarity function, indexes the similarity between data points. Gilboa et al. (2006) developed the theory for the case where y is a real number, while Billot et al. (2005) developed the theory for the case where y is a probability vector. Gilboa et al. (2011) and Lieberman (2012) extended the theory to the problems of density estimation and autoregression, respectively.

Both Gilboa et al. (2006) and Billot et al. (2005) provided axiomatizations of empirical similarity theory with a generic similarity function. Although neither assumed a particular similarity function or even a particular functional form, Gilboa et al. (2006) expressed interest in similarity functions that are based on a metric. Two similarity functions have received the most attention in the literature: the reciprocal similarity function, $s(x_i, x) = 1/(1 + d(x_i, x))$, and the exponential similarity function, $s(x_i, x) = \exp(-d(x_i, x))$, where d is a metric on $\mathbb{R}^n$.

Billot et al. (2008) provided an axiomatization of an exponential similarity function that is based on a norm, i.e., $s(x_i, x) = \exp(-N(x - x_i))$ for some norm N on $\mathbb{R}^n$. They also axiomatized the special cases of an exponential similarity function based on the standard Euclidean norm and on a weighted Euclidean norm. Recently, the literature has focused on models with an exponential similarity function based on a weighted Euclidean metric (Lieberman, 2010; Gilboa et al., 2011).[2]

Defining similarity in terms of a metric (whether or not the metric is induced by a norm) imposes restrictions on the similarity function that follow from the properties of a metric.[3] In the case of the exponential similarity function, the symmetry of d implies that s is symmetric, i.e., $s(x_i, x) = s(x, x_i)$, while the triangle inequality for d implies that s satisfies a form of multiplicative transitivity, namely, $s(x_i, x) \ge s(x_i, z)s(z, x)$ (Billot et al., 2008). In many applications of empirical similarity theory, the restrictions on the similarity function that follow from defining similarity in terms of a metric may be reasonable. In other applications, however, these restrictions are problematic.[4]

[*] Tel.: +1 2026616589.
E-mail address: jct48@law.georgetown.edu.

[1] The theory is closely related to case-based decision theory (Gilboa and Schmeidler, 2001).

[2] A notable exception is Gayer et al. (2007), which estimated an empirical similarity model with a reciprocal similarity function based on a weighted Euclidean metric.

[3] Note, however, that the similarity function in empirical similarity theory satisfies positivity irrespective of whether it is based on a metric. The positivity of the similarity function follows from the axiomatizations in Gilboa et al. (2006) and Billot et al. (2005).

[4] Indeed, some argue that these restrictions are problematic in any application that purports to model human reasoning. There is work in psychology that questions whether human similarity judgments obey metric properties such as symmetry and the triangle inequality (e.g., Tversky, 1977; Tversky and Gati, 1982).

0165-4896/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.mathsocsci.2013.07.005

In particular, there are numerous applications of empirical similarity theory in which the apposite notion of similarity is not symmetric. Consider, for example, applying the theory to model judicial decisions in legal cases. In a legal model, because the similarity function s determines how much weight the judge gives the outcome of a prior case in deciding the outcome of the new case, one properly interprets s as measuring precedential influence. But precedential influence generally is not symmetric. In a legal system with hierarchical courts, precedential influence depends not only on fact similarity (distance in "fact space") but also on precedential authority (relative position in the judicial hierarchy). Fact similarity is symmetric, but precedential authority is not symmetric. All else equal, the precedential authority of a case decided by a superior court is greater than the precedential authority of a case decided by an inferior court. Therefore, if the prior case was decided by a superior court, its influence on the outcome of the new case ought to be greater than the influence of the new case on the outcome of the prior case (under the counterfactual that the new case was decided before the prior case).

This paper suggests a similarity function for applications of empirical similarity theory in which the notion of similarity is not necessarily symmetric. More specifically, I propose defining similarity in terms of a *quasimetric*, i.e., a function that satisfies all the properties of a metric except for symmetry (Wilson, 1931). Quasimetrics and other asymmetric distance measures have been used in operations research and related fields to model, inter alia, rush-hour traffic, flight in the presence of wind, marine navigation in the presence of currents, and transportation on sloped terrain (Drezner and Wesolowsky, 1989). In the empirical similarity literature, Lieberman (2012) has used a different asymmetric similarity function (not based on a quasimetric) in an application of his model of similarity-based autoregression.

The paper is organized as follows. Section 2 describes the asymmetric empirical similarity model and proposes an asymmetric distance measure on which to base the similarity function. It then explores the properties of the model and provides motivation for the proposed measure. In Section 3, I prove that the proposed measure is a quasimetric. Section 4 situates the proposed measure in the class of quasimetrics induced by *skewed norms* (Plastria, 1992) and discusses the virtues of this class. Concluding remarks appear in Section 5. In the Appendix, I provide a "skewness" axiom that, when imposed in lieu of the "symmetry" axiom in the main result of Billot et al. (2008), characterizes an exponential similarity function based on a skewed norm.

2. Asymmetric empirical similarity model

2.1. The model

Following Gilboa et al. (2006) and Billot et al. (2005), I assume that, given a new data point $x \in \mathbb{R}^n$, a database of prior cases $C = (x_i, y_i)_{i \leq t}$, $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ for all $i \leq t$, and a similarity function $s : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{++}$, one assesses the value of a real variable y according to the formula

$$y = Y(C, x) = \frac{\sum\limits_{i \leq t} s(x_i, x) y_i}{\sum\limits_{i \leq t} s(x_i, x)}, \tag{1}$$

where Y is defined on the set of all databases, $\mathcal{C} = \bigcup_{t \geq 1} \left(\mathbb{R}^{n+1}\right)^t$, and for all data points $x \in \mathbb{R}^n$. In what follows, I refer to data points as "inputs" and to assessments as "outcomes".

According to Eq. (1), the outcome y in the new case is a weighted average of the outcomes $y_1, \ldots, y_t$ in the prior cases. The weight placed on a prior outcome y_i in the determination of the new outcome y depends on the degree to which the input x_i in the prior case is similar to the input x in the new case. The degree of similarity is given by s. The greater is the similarity of a prior input x_i to the new input x, the greater is the weight given to the prior outcome y_i in the determination of the new outcome y. Thus, I interpret s as measuring the "influence" of a prior case on the assessment of the new case.

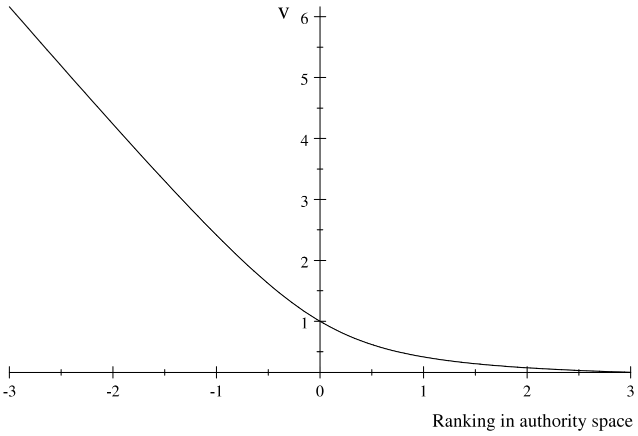

Fig. 1. The proportion υ as a function of ranking in authority space.

Departing from the prior literature, I define the input space as the Cartesian product of $\mathcal{A} = \mathbb{R}^{n-1}$ and $\mathcal{B} = \mathbb{R}$, where $\mathcal{A}$ is the multidimensional space on which inputs are substantively compared (the "comparison space") and $\mathcal{B}$ is the unidimensional space on which they are ranked in terms of authority (the "authority space"). Accordingly, each input $x = (x^a, x^b) \in \mathbb{R}^n$ comprises an a-component $x^a \in \mathcal{A}$ and a b-component $x^b \in \mathcal{B}$. In a legal model, for instance, x^a gives the position in fact space (the comparison space) and x^b gives the position in the judicial hierarchy (the authority space).

In accordance with the prior literature, I assume that the similarity of a prior input x_i to a new input x is a decreasing function of the distance μ in input space from x_i to x. That is, I assume

$$s(x_i, x) = f(\mu(x_i, x)) \tag{2}$$

for some decreasing function $f : \mathbb{R}_+ \to \mathbb{R}_{++}$. In addition, I assume $\mu(x, x) = 0$ and $f(0) = 1$, which together imply that similarity is reflexive—i.e., $s(x, x) = 1$. However, I do not assume that similarity is symmetric. That is, I do not assume $s(x_i, x) = s(x, x_i)$ for $x_i \neq x$. Rather, I wish to allow that similarity is asymmetric—i.e., $s(x_i, x) \neq s(x, x_i)$ for $x_i \neq x$.

To this end, I suggest defining μ as follows. For all $x_i, x \in \mathbb{R}^n$, $x_i \neq x$, let

$$\mu(x_i, x) = \upsilon(\theta_i) d_w(x_i^a, x^a), \tag{3}$$

where d_w denotes the weighted Euclidean metric, i.e.,

$d_w(x_i^a, x^a) = \sqrt{\sum_{j=1}^{n-1} w_j (x_j^a - x_{ij}^a)^2}$, $w_j > 0$ for all j,

$\upsilon(\theta_i) = \sec\theta_i + \tan\theta_i$,

and θ_i is the polar angle of $(d_w(x_i^a, x^a), \sqrt{w_n}(x^b - x_i^b))$, $w_n > 0$. In other words, I suggest defining the distance μ from the input x_i of a prior case to the input x of a new case as a proportion υ of the weighted distance d_w between them in comparison space. All else constant, the proportion υ decreases (increases) as the cardinal ranking in authority space of the prior case increases (decreases) relative to the new case (see Fig. 1). Consequently, μ is asymmetric—$\mu(x_i, x) \neq \mu(x, x_i)$ for $x_i \neq x$. Because f is decreasing, the asymmetry of μ implies that s is asymmetric as well.

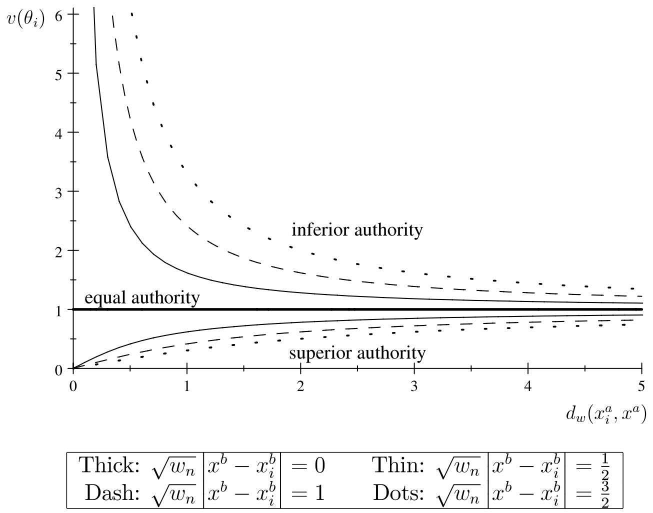

Fig. 2. Visualization of v.

In Section 3, I prove that μ is a quasimetric on $\mathbb{R}^n$. That is, I prove that μ satisfies all the properties of metric, apart from symmetry.[5] In the remainder of this section, I explore the properties of the model and provide motivation for the specification of μ.

2.2. *Properties of the model*

As noted in Section 2.1, the model defines the distance μ from the input x_i of a prior case to the input x of a new case as a proportion v of the weighted distance d_w between them in comparison space. As Fig. 2 illustrates, $v < 1$ if the prior case ranks above the new case in authority space ($x_i^b > x^b$), $v = 1$ if the prior case and the new case have the same rank in authority space ($x_i^b = x^b$), and $v > 1$ if the prior case ranks below the new case in authority space ($x_i^b < x^b$). I refer to $|1 - v|$ as the *authority factor*. All else constant, the authority factor is positively related to the weighted distance in authority space, $\sqrt{w_n}\left|x^b - x_i^b\right|$, and negatively related to the weighted distance in comparison space, d_w. In other words, the authority factor is greater (i) the greater is the authority gap between the prior case and the new case, (ii) the more comparable is the prior case to the new case, and (iii) the greater is the weight placed on authority relative to comparability.

Fig. 3 displays the relationship in the model between influence (s), comparability (d_w), and authority (v). It assumes an exponential similarity function, $s(x_i, x) = \exp(-\mu(x_i, x))$, implying that influence decays exponentially with input distance. As the figure illustrates, the influence of a prior case on the assessment of the new case is greatest when the prior case is perfectly comparable to the new case ($x_i^a = x^a \Leftrightarrow d_w = 0$), and it decays at rate v as comparability diminishes (i.e., as d_w increases). Both the influence at $d_w = 0$ and the rate of decay for $d_w > 0$ differ depending on the authority of the prior case. If the prior case has superior authority ($x_i^b > x^b$), the influence at $d_w = 0$ is the highest possible ($s = 1$) and the rate of decay for $d_w > 0$ is lower ($v < 1$). If, however, the prior case has inferior authority ($x_i^b < x^b$), the influence at $d_w = 0$ is lower ($s < 1$) and the rate of decay for $d > 0$ is higher ($v > 1$). All else equal, the influence of a prior case with superior authority is greater than the influence of a prior case with inferior authority. Moreover, this influence differential increases with the authority gap, $\sqrt{w_n}\left|x^b - x_i^b\right|$, as well as the degree of comparability, $1/d_w$.

2.3. *Motivation for μ*

We can motivate the specification of μ in the model by analogy to the problem of measuring the work required to slide a block along an inclined plane (see, e.g., Hodgson et al. (1987)). Let $x_i = (x_i^a, x_i^b) \in \mathbb{R}^3$ denote the block's origination point and $x = (x^a, x^b) \in \mathbb{R}^3$ denote its destination point, where $x_i^a, x^a \in \mathcal{A} = \mathbb{R}^2$ and $x_i^b, x^b \in \mathcal{B} = \mathbb{R}$. Assume distance is measured by the weighted Euclidean metric, d_w. The distance along the plane from x_i to x is

$$r_i = d_w(x_i, x) = \|x - x_i\|_w ,$$

where $\|\cdot\|_w$ denotes the weighted Euclidean norm. The angle of incline from x_i to x is

$$\theta_i = \arctan\left(\frac{\sqrt{w_n}(x^b - x_i^b)}{d_w(x_i^a, x^a)}\right).$$

The vertical distance (height) from x_i to x is

$$h_i = r_i \sin\theta_i = \sqrt{w_n}(x^b - x_i^b),$$

which follows from the fact that (r_i, θ_i) are the polar coordinates of $(d_w(x_i^a, x^a), \sqrt{w_n}(x^b - x_i^b))$.

The work required to slide the block from x_i to x is the product of the force required to slide the block and the distance over which the force is exerted. The force required to slide the block has two components—the force required to overcome friction, F, and the force required to overcome gravity, which equals the weight of the block, W. Thus, the work required to slide the block from x_i to x is

$$\begin{aligned} K(x_i, x) &= Fr_i + Wh_i \\ &= F\,\|x - x_i\|_w + W\sqrt{w_n}(x^b - x_i^b). \end{aligned}$$

Observe that K is asymmetric—$K(x_i, x) \neq K(x, x_i)$ for $x_i \neq x$. The work required to slide the block uphill is greater than the work required to slide the block the same distance down the same hill.

Turning from the physical world to the world of analogical reasoning, let us interpret K as measuring the work that is required to reason by analogy from a prior case at x_i to a new case at x. All else constant, the work required is greater (i) the less comparable is the prior case (i.e., the greater is $d_w(x_i^a, x^a)$) and (ii) the less authoritative is the prior case (i.e., the greater is $x^b - x_i^b$). Under this interpretation, F and W are cost parameters that determine the marginal work associated with decreases in comparability and authority. More specifically, F determines the marginal work associated with a decrease in comparability, and F and W jointly determine the marginal work associated with a decrease in authority. Observe that if we normalize $F = W = 1$,[6] then K corresponds to μ:

$$\begin{aligned} \mu(x_i, x) &= v(\theta_i) d_w(x_i^a, x^a) \\ &= (\sec\theta_i + \tan\theta_i)(r_i \cos\theta_i) \\ &= r_i + r_i \sin\theta_i \\ &= \|x - x_i\|_w + \sqrt{w_n}(x^b - x_i^b) = K(x_i, x). \end{aligned}$$

Defining similarity in terms of μ, therefore, may be motivated by the intuition that the influence of a prior case on the assessment of a new case is negatively related to the work required to draw an analogy from the prior case to the new case. In other words, the more strained is the analogy, the less influential is the prior case.

3. Proof that μ is a quasimetric

In this section, I prove that the function μ defined in Section 2.1 is a quasimetric on $\mathbb{R}^n$. Recall the definition of a quasimetric:

[5] By contrast, the asymmetric distance measure suggested by Lieberman (2012) in the context of similarity-based autoregression is not a quasimetric—it does not always satisfy the triangle inequality on $\mathbb{R}^n$ for $n > 1$.

[6] In the physical world, $F = W \sin\theta_i$. That is, the force of friction depends on the weight of the block and the angle of incline of the plane, and $F = W$ only if $\theta_i = \pi/2$ (45°). In the world of analogical reasoning, however, there is no reason why this relationship must or even should hold, or indeed why F must or even should depend at all on W or θ_i. In this world, therefore, we can have $F = W = 1$.

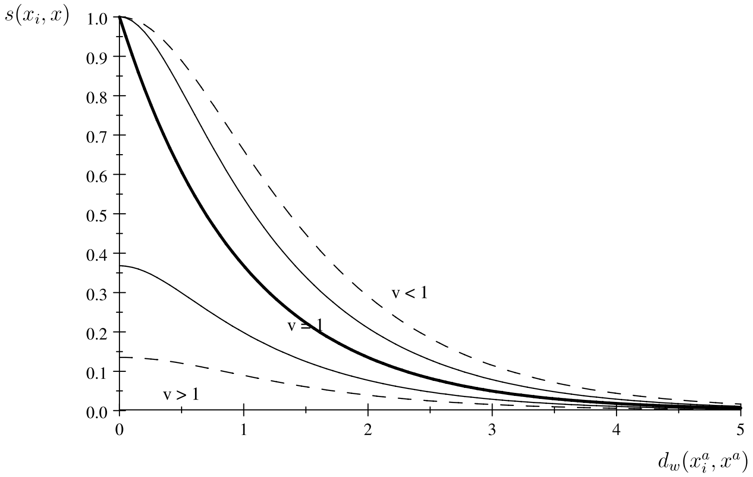

Fig. 3. Relationship between influence (s), comparability (d_w), and authority (v).

Definition 1. A function $q : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a *quasimetric* on $\mathbb{R}^n$ if for all $x_i, x \in \mathbb{R}^n$:

(i) $q(x_i, x) \geq 0$;
(ii) $q(x_i, x) = 0$ if and only if $x_i = x$; and
(iii) $q(x_i, x) \leq q(x_i, z) + q(z, x)$ for any $z \in \mathbb{R}^n$ (triangle inequality).

Note that a quasimetric is a metric if it also satisfies symmetry: $q(x_i, x) = q(x, x_i)$. A quasimetric is not necessarily symmetric, i.e., in general $q(x_i, x) \neq q(x, x_i)$.

Proposition 1. *For all $x_i^a, x^a \in \mathbb{R}^{n-1}$ and $x_i^b, x^b \in \mathbb{R}$, with $x_i = (x_i^a, x_i^b)$ and $x = (x^a, x^b)$, let $\mu(x_i, x) = v(\theta_i) d_w(x_i^a, x^a)$, where d_w denotes the weighted Euclidean distance on $\mathbb{R}^{n-1}$, $v(\theta_i) = \sec\theta_i + \tan\theta_i$, and θ_i is the polar angle of $(d_w(x_i^a, x^a), \sqrt{w_n}(x^b - x_i^b))$, $w_n > 0$. Then μ is a quasimetric on $\mathbb{R}^n$.*

Proof. Recall from Section 2.3 that

$$\mu(x_i, x) = v(\theta_i) d_w(x_i^a, x^a) = \|x - x_i\|_w + \sqrt{w_n}(x^b - x_i^b).$$

(i) Observe that $\|x - x_i\|_w = \sqrt{(d_w(x_i^a, x^a))^2 + w_n(x^b - x_i^b)^2}$. It follows that $\|x - x_i\|_w \geq |\sqrt{w_n}(x^b - x_i^b)|$, and hence $\mu(x_i, x) \geq 0$.

(ii) If $x_i = x$ then $\|x - x_i\|_w = 0$ and $(x^b - x_i^b) = 0$, and hence $\mu(x_i, x) = 0$. Now suppose $\mu(x_i, x) = 0$ but $x_i \neq x$. If $x_i \neq x$ then $\|x - x_i\|_w > 0$ and $\sqrt{w_n}(x^b - x_i^b) \neq 0$. However, because $\|x - x_i\|_w \geq |\sqrt{w_n}(x^b - x_i^b)|$, this implies $\mu(x_i, x) > 0$, which contradicts $\mu(x_i, x) = 0$.

(iii) Take any $z \in \mathbb{R}^n$. To prove that μ satisfies the triangle inequality, we must show that $\mu(x_i, x) \leq \mu(x_i, z) + \mu(z, x)$. The condition holds if and only if

$$\begin{aligned}\|x - x_i\|_w \leq\ & \|x - z\|_w + \|z - x_i\|_w + \sqrt{w_n}(z^b - x_i^b) \\ & + \sqrt{w_n}(x^b - z^b) - \sqrt{w_n}(x^b - x_i^b).\end{aligned}$$

Observe that $\sqrt{w_n}(z^b - x_i^b) + \sqrt{w_n}(x^b - z^b) - \sqrt{w_n}(x^b - x_i^b) = 0$. Observe further that

$$\|x - x_i\|_w = \|(x - z) + (z - x_i)\|_w \leq \|x - z\|_w + \|z - x_i\|_w$$

by the subadditivity of norms. Hence, the condition holds. ■

4. Skewed norms

The quasimetric μ defined in Section 2.1 belongs to the class of quasimetrics induced by skewed norms. Plastria (1992) defines this class as follows. Let N be a norm on $\mathbb{R}^n$ and $p \in \mathbb{R}^n$. Define the *linearly perturbed norm function* $L(N, p)$ by

$$L(N, p)(x) = N(x) - \langle p, x \rangle\,,$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product. Observe that $L(N, p)$ is positively homogeneous, $L(N, p)(\lambda x) = \lambda L(N, p)(x)$ for $\lambda > 0$, and subadditive, $L(N, p)(x + z) \leq L(N, p)(x) + L(N, p)(z)$. If $L(N, p)$ is also positive definite, $L(N, p)(x) > 0$ for all $x \neq \mathbf{0}$, then it is a skewed norm. One can readily show that if $L(N, p)$ is a skewed norm then $q(x_i, x) = L(N, p)(x - x_i)$ is a quasimetric.

Recall from Section 2.3 that we can rewrite μ as

$$\mu(x_i, x) = \|x - x_i\|_w + \sqrt{w_n}(x^b - x_i^b), \tag{4}$$

where $\|\cdot\|_w$ denotes the weighted Euclidean norm. From (4), we can readily see that μ is induced by a linearly perturbed norm $L(N_\mu, p_\mu)$, where $N_\mu = \|\cdot\|_w$ and $p_\mu = (0, \ldots, 0, -\sqrt{w_n})$. Moreover, we can readily show that $L(N_\mu, p_\mu)$ is positive definite, and therefore qualifies as a skewed norm. Thus, μ is a quasimetric induced by a skewed norm.

A number of other quasimetrics in this class have been developed for use in continuous location problems in operations research and related fields; see, e.g., Hodgson et al. (1987) (p-centroid location), Drezner and Wesolowsky (1989) (single facility location), and Cera et al. (2008) (hunter location). The importance and popularity of quasimetrics induced by skewed norms stem in part from two facts. First, they are closely and simply related to the familiar class of metrics induced by norms. Second, they are convex functions, and thus amenable to the powerful methods of convex analysis and optimization (Plastria, 2009).

Two additional virtues of this class are its breadth and flexibility. Although μ may be an attractive specification for many applications because it is characterized by a familiar norm, $N_\mu = \|\cdot\|_w$, and a one-parameter perturbation vector, $p_\mu = (0, \ldots, 0, -\sqrt{w_n})$, one can readily fashion an alternative specification by selecting a different norm N or perturbation vector p. Provided that $L(N, p)$ is positive definite,[7] the alternative specification will be a quasimetric, and hence provide asymmetry while preserving the triangle inequality.

5. Concluding remarks

The standard empirical similarity model defines similarity in terms of a metric. This imposes restrictions on the similarity function that follow from the properties of a metric, including symmetry. For many applications of empirical similarity theory, however, the apposite notion of similarity is not symmetric. Defining similarity in terms of a quasimetric allows the model to capture

[7] Note that $L(N, p)$ is positive definite if $N^d(p) < 1$, where N^d denotes the dual of N (Plastria, 1992).

situations in which similarity judgments are asymmetric. Such situations arise, for example, when the influence of a prior case depends not only on its comparability but also on its relative weight of authority. A prime example is judicial decision making in a legal system with hierarchical courts.

The asymmetric similarity function proposed in this paper is based on a quasimetric induced by a skewed norm. The class of quasimetrics induced by skewed norms has several virtues, including a close resemblance to the familiar class of metrics induced by norms. In the Appendix, I provide a "skewness" axiom that, when imposed in lieu of the "symmetry" axiom in the main result of Billot et al. (2008), characterizes an exponential similarity function based on a skewed norm. The skewness axiom essentially postulates exponential discounting of the influence of prior cases with inferior authority relative to equidistant prior cases with superior authority. It thus illuminates an intuitive and observable implication of the exponential empirical similarity model when similarity is based on a skewed norm.

Appendix

As noted in Section 1, Billot et al. (2008) provided an axiomatization of an exponential similarity function that is based on a norm, i.e.,

$$s(x_i, x) = \exp(-N(x - x_i)) \tag{5}$$

for some norm N on $\mathbb{R}^n$. More specifically, they assumed the data generating process was described by Eq. (1) and imposed five axioms on Y that characterized (5). The five axioms are shift invariance, ray monotonicity, symmetry, ray shift invariance, and self-relevance.

The symmetry axiom requires that for every $x \in \mathbb{R}^n$,

$$Y(((\mathbf{0}, 1), (x, 0)), x) = Y(((x, 1), (\mathbf{0}, 0)), \mathbf{0}). \tag{6}$$

Eq. (6) considers two situations. In the first, outcome 1 occurred at $\mathbf{0} \in \mathbb{R}^n$ and outcome 0 occurred at $x \in \mathbb{R}^n$, and an assessment is requested for $x \in \mathbb{R}^n$. In the second, outcome 1 occurred at $x \in \mathbb{R}^n$ and outcome 0 occurred at $\mathbf{0} \in \mathbb{R}^n$, and an assessment is requested for $\mathbf{0} \in \mathbb{R}^n$. Symmetry requires that the assessment is the same in both situations.

In the presence of shift invariance,[8] the symmetry axiom is equivalent to requiring that for every $x \in \mathbb{R}^n$,

$$Y(((\mathbf{0}, 1), (2x, 0)), x) = Y(((2x, 1), (\mathbf{0}, 0)), x). \tag{7}$$

To see this, observe that (6) is equivalent to

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

which, because $s(\mathbf{0}, \mathbf{0}) = s(x, x) = 1$, holds if and only if $s(\mathbf{0}, x) = s(x, \mathbf{0})$. Note that $s(\mathbf{0}, x) = s(x, \mathbf{0})$ if and only if

$$\frac{s(\mathbf{0}, \mathbf{0})}{s(\mathbf{0}, \mathbf{0}) + s(x, \mathbf{0})} = \frac{s(x, x)}{s(x, x) + s(\mathbf{0}, x)}.$$

It follows that (6) is equivalent to

$$Y(((\mathbf{0}, 1), (x, 0)), \mathbf{0}) = Y(((x, 1), (\mathbf{0}, 0)), x). \tag{6$'$}$$

By the shift invariance axiom,

$$Y(((\mathbf{0}, 1), (x, 0)), \mathbf{0}) = Y(((x, 1), (2x, 0)), x).$$

It follows that (6′) is equivalent to

$$Y(((x, 1), (2x, 0)), x) = Y(((x, 1), (\mathbf{0}, 0)), x). \tag{6$''$}$$

Observe that (6″) is equivalent to

$$\frac{s(x, x)}{s(x, x) + s(2x, x)} = \frac{s(x, x)}{s(x, x) + s(\mathbf{0}, x)},$$

which holds if and only if $s(\mathbf{0}, x) = s(2x, x)$. Note that $s(\mathbf{0}, x) = s(2x, x)$ if and only if

$$\frac{s(\mathbf{0}, x)}{s(\mathbf{0}, x) + s(2x, x)} = \frac{s(2x, x)}{s(2x, x) + s(\mathbf{0}, x)},$$

which is equivalent to (7).

Much like (6), Eq. (7) considers two situations: (i) outcome 1 occurred at $\mathbf{0} \in \mathbb{R}^n$ and outcome 0 occurred at $2x \in \mathbb{R}^n$ and (ii) outcome 1 occurred at $2x \in \mathbb{R}^n$ and outcome 0 occurred at $\mathbf{0} \in \mathbb{R}^n$. In both situations, an assessment is requested at $x \in \mathbb{R}^n$. Symmetry requires that the assessments are equal. Stated another way, symmetry requires that the log ratio of the assessments is identically zero:

$$\ln\left(\frac{Y(((\mathbf{0}, 1), (2x, 0)), x)}{Y(((2x, 1), (\mathbf{0}, 0)), x)}\right) = 0.$$

Observe that (7) is equivalent to $s(\mathbf{0}, x) = s(2x, x)$. Intuitively, therefore, symmetry requires that, all else equal, the influence of a prior case on the assessment of a new case at x is the same whether the prior case occurs at $\mathbf{0}$ or $2x$.

The central claim of this Appendix is that the following "skewness" axiom, when imposed in lieu of the symmetry axiom in the main result of Billot et al. (2008), characterizes an exponential similarity function based on a skewed norm, i.e., $s(x_i, x) = \exp(-\gamma(x - x_i))$, where $\gamma(x) = N(x) - \langle p, x\rangle$ for some norm N on $\mathbb{R}^n$ and $p \in \mathbb{R}^n$ and $\gamma(x) > 0$ for all $x \neq \mathbf{0}$.

Axiom 1 (*Skewness*). For every $x \in \mathbb{R}^n$,

$$Y(((\mathbf{0}, 1), (2x, 0)), x) = Y(((2x, 1), (\mathbf{0}, 0)), x) \times \exp(-2\langle p, x\rangle) \tag{8}$$

for some $p \in \mathbb{R}^n$.

Exactly like (7), (8) requests an assessment at $x \in \mathbb{R}^n$ in situations (i) and (ii). In contrast to symmetry, however, skewness does not require that the assessment is the same in both situations. Instead, skewness requires that the assessment in the prior situation equals the assessment in the latter situation discounted by a factor of $\exp(-2\langle p, x\rangle)$. Stated another way, skewness requires that the log ratio of the assessments is linear:

$$\ln\left(\frac{Y(((\mathbf{0}, 1), (2x, 0)), x)}{Y(((2x, 1), (\mathbf{0}, 0)), x)}\right) = -2\langle p, x\rangle.$$

Intuitively, skewness requires that, all else equal, the influence of a prior case on the assessment of a new case at x is discounted by a factor of $\exp(-2\langle p, x\rangle)$ when the prior case occurs at $\mathbf{0}$ relative to when it occurs at $2x$. Observe that the discount factor equals one for $p = \mathbf{0}$, in which case the assessments are the same and skewness reduces to symmetry.

The claim may be established in three steps. The first step is to prove that in the absence of the symmetry axiom, the remaining four axioms (shift invariance, ray monotonicity, ray shift invariance, and self-relevance) characterize an exponential similarity function that is based on a *gauge*, i.e., $s(x_i, x) = \exp(-g(x - x_i))$ for some gauge g on $\mathbb{R}^n$.[9] The proof comprises a straightforward modification of the proof of Theorem 1 in Billot et al. (2008); hence, it is omitted.

[8] The shift invariance axiom requires that for every database $C = (x_i, y_i)_{i \le t} \in \mathcal{C}$ and every $x, z \in \mathbb{R}^n$, $Y((x_i+z, y_i)_{i \le t}, x+z) = Y((x_i, y_i)_{i \le t}, x)$. See Billot et al. (2008).

[9] A gauge is a nonnegative function $g : \mathbb{R}^n \to \mathbb{R}_+$ satisfying:

(i) $g(x) = 0$ if and only if $x = \mathbf{0}$;
(ii) $g(\lambda x) = \lambda g(x)$ for all $x \in \mathbb{R}^n$ and $\lambda \ge 0$; and
(iii) $g(x, z) \le g(x) + g(z)$ for all $x, z \in \mathbb{R}^n$.

See, e.g., Rockafellar (1970).

The second step is to show that Axiom 1 holds if and only if $g(-x) - g(x)$ is linear. Observe that Axiom 1 is equivalent to

$$\frac{s(\mathbf{0}, x)}{s(\mathbf{0}, x) + s(2x, x)} = \frac{s(2x, x)}{s(2x, x) + s(\mathbf{0}, x)} \times \exp(-2\langle p, x\rangle),$$

which holds if and only if $s(\mathbf{0}, x) = s(2x, x) \times \exp(-2\,\langle p, x\rangle)$. From $s(x_i, x) = \exp(-g(x - x_i))$, it follows that Axiom 1 holds if and only if $\exp(-g(x)) = \exp(-g(-x)) \times \exp(-2\,\langle p, x\rangle)$, which in turn holds if and only if $g(-x) - g(x) = -2\,\langle p, x\rangle$.

The final step is to invoke Theorem 3 of Plastria (1992), which establishes, inter alia, that $g(-x) - g(x)$ is linear if and only if g is a skewed norm.

Assuming that the assessments Y are observable, Axiom 1 may be interpreted as an observable implication of the exponential empirical similarity model when similarity is based on a skewed norm. In principle, therefore, Axiom 1 is testable. A difficulty in testing Axiom 1 is that it requires the existence of an n-parameter vector p such that (8) holds for every x. The difficulty is lessened somewhat in the case of μ, however, because the p associated with μ has only one free parameter, w_n.[10]

References

Billot, A., Gilboa, I., Samet, D., Schmeidler, D., 2005. Probabilities as similarity-weighted frequencies. Econometrica 73, 1125–1136.

Billot, A., Gilboa, I., Schmeidler, D., 2008. Axiomatization of an exponential similarity function. Mathematical Social Sciences 55, 107–115.

Cera, M., Mesa, J.A., Ortega, F.A., Plastria, F., 2008. Locating a central hunter on the plane. Journal of Optimization Theory and Applications 136, 155–166.

Drezner, Z., Wesolowsky, G.O., 1989. The asymmetric distance location problem. Transportation Science 23, 201–207.

Gayer, G., Gilboa, I., Lieberman, O., 2007. Rule-based and case-based reasoning in housing prices. B.E. Journal of Theoretical Economics 7 (1) (Advances). Article 10.

Gilboa, I., Lieberman, O., Schmeidler, D., 2006. Empirical similarity. The Review of Economics and Statistics 88, 433–444.

Gilboa, I., Lieberman, O., Schmeidler, D., 2011. A similarity-based approach to prediction. Journal of Econometrics 162, 124–131.

Gilboa, I., Schmeidler, D., 2001. A Theory of Case-Based Decisions. Cambridge University Press, Cambridge.

Hodgson, M.J., Wong, R.T., Honsaker, J., 1987. The p-centroid problem on an inclined plane. Operations Research 35, 221–233.

Lieberman, O., 2010. Asymptotic theory for empirical similarity models. Econometric Theory 26, 1032–1059.

Lieberman, O., 2012. A similarity-based approach to time-varying coefficient non-stationary autoregression. Journal of Time Series Analysis 33, 484–502.

Plastria, F., 1992. On destination optimality in asymmetric distance Fermat–Weber problems. Annals of Operations Research 40, 355–369.

Plastria, F., 2009. Asymmetric distances, semidirected networks and majority in Fermat–Weber problems. Annals of Operations Research 167, 121–155.

Rockafellar, R.T., 1970. Convex Analysis. Princeton University Press, Princeton, NJ.

Tversky, A., 1977. Features of similarity. Psychological Review 84, 327–352.

Tversky, A., Gati, I., 1982. Similarity, separability, and the triangle inequality. Psychological Review 89, 123–154.

Wilson, W.A., 1931. On quasi-metric spaces. American Journal of Mathematics 53, 675–684.

[10] Recall that μ is induced by the skewed norm $L(N_\mu, p_\mu)$, where $N_\mu = \|\cdot\|_w$ and $p_\mu = (0, \ldots, 0, -\sqrt{w_n})$.