AXIOMATIZATION OF AN EXPONENTIAL SIMILARITY FUNCTION

By

Antoine Billot, Itzhak Gilboa and David Schmeidler

October 2004

COWLES FOUNDATION DISCUSSION PAPER NO. 1485

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY

Box 208281

New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Axiomatization of an Exponential Similarity Function

Antoine Billot,[*] Itzhak Gilboa,[†] and David Schmeidler[‡]

Preliminary Draft – March 2004

Abstract

An agent is asked to assess a real-valued variable y based on certain characteristics $x = (x^1, ..., x^m)$, and on a database consisting of n observations of $(x^1, ..., x^m, y)$. 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 $x^1_{n+1}, ..., x^m_{n+1}$, associated with y_{n+1}, and the previously observed vector, $x^1_i, ..., x^m_i$. 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.

1 Introduction

Consider a problem in which one attempts to assess the value of a variable y based on the values of relevant variables, $x = (x^1, ..., x^m)$, and on a database consisting of past observations of the variables $(x_i, y_i) = (x^1_i, ..., x^m_i, y_i)$, $i = 1, ..., n$. One approach to deal with this classical 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

[*]University de Paris II, IUF, and CERAS-ENPC. billot@u-paris2.fr

[†]Tel-Aviv University and Yale University. igilboa@post.tau.ac.il

[‡]Tel-Aviv University and The Ohio State University. schmeid@post.tau.ac.il

a database of past observations, $B = (x_i, y_i)_{i \leq n}$, where $x_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$, and a new data point $x_{n+1} \in \mathbb{R}^m$, generate the prediction

$$y^s_{n+1} = \frac{\sum_{i \leq n} s(x_i, x_{n+1}) y_i}{\sum_{i \leq n} s(x_i, x_{n+1})}$$

This formula was suggested and axiomatized in Gilboa, Lieberman, and Schmeidler (2004).[1] They assume that, given any database B and any new data point $x_{n+1} \in \mathbb{R}^m$, a predictor has an ordering over $\mathbb{R}$, interpreted as "more likely than". They show that this ordering satisfies certain axioms if and only there exists a similarity function such that the ordering ranks possible predictions y according to their proximity to y^s_{n+1}.

In this paper we consider a prediction function $Y(B, x)$ that, for every database B (consisting of $n \geq 1$ observations) and every new point $x \in \mathbb{R}^m$, generates a real-valued prediction. We interpret Y as a maximizer of the "more likely than" relation, and we will therefore assume that there exists a similarity function s for which $Y((x_i, y_i)_{i \leq n}, x_{n+1}) = y^s_{n+1}$. Our interest is in the relationship between properties of this function Y and the function s. We provide axiom, stated in terms of the function Y, that are equivalent to the statement that the similarity function $s(x, z)$ takes the form $s(x, z) = \exp[-n(x - z)]$ for some norm n.

The next section provides the model and the main result. It is followed by comments on several special case of the norm n, and a general discussion. Proofs are to be found in an appendix.

2 Result

Let $\mathbb{B} = \cup_{n \geq 1} (\mathbb{R}^{k+1})^n$ be the set of possible databases. A database $B \in \mathbb{B}$ is a vector of n observations of $k + 1$ real numbers. It will be written as

[1] It is reminiscent of derivations in Gilboa and Schmeidler (2003) and in Billot, Gilboa, Samet, and Schmeidler (2004). 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.

$B = (x_i, y_i)_{i \leq n}$, where $x_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$.

Let $Y : \mathbb{B} \times \mathbb{R}^k \to \mathbb{R}$ be a function that predicts the value $Y(B, x)$ given a database $B = (x_i, y_i)_{i \leq n}$ and a new point $x = x_{n+1} \in \mathbb{R}^m$. We assume that there exists a similarity function $s : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_{++}$ such that, for every $B = (x_i, y_i)_{i \leq n}$ and every $x \in \mathbb{R}^m$,

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

Gilboa, Lieberman, and Schmeidler (2004) provide an axiomatization of this formula. Their derivation is done for each x separately. That is, they fix $x \in \mathbb{R}^m$ and consider the rankings of possible values of $Y(B, x)$ for various databases B. The function that they obtain is, therefore, $s(\cdot, x)$ for each x. This function 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 impose the following axioms on Y:

A1 Shift: For every $B = (x_i, y_i)_{i \leq n} \in \mathbb{B}$, and every $x, \delta \in \mathbb{R}^m$, $Y((x_i + \delta, y_i)_{i \leq n}, x + \delta) = Y((x_i, y_i)_{i \leq n}, x)$.

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 points in the database, accompanied by the same shift in the new point for which prediction is required, will not affect the predicted value Y.

The following 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 $x = 0$.

A2 Ray Monotonicity: For every $x, z \in \mathbb{R}^m$, $Y(((\lambda x, 1), (z, -1)), 0)$ is strictly decreasing in $\lambda \geq 0$.

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, equation (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 0. A2 states that, if we move λx further away from 0 (along the same vector), it will be considered less similar to 0, and the prediction Y will decrease.

A3 Symmetry: For every $x \in \mathbb{R}^m$, $Y(((x, 1), (0, 0)), 0) = Y(((0, 1), (x, 0)), x)$.

A3 considers two situations. In the first, one has observed the value 1 for $x \in \mathbb{R}^m$, and the value 0 for $0 \in \mathbb{R}^m$, and one is asked to make a prediction for $0 \in \mathbb{R}^m$. In the second situation, the roles are reversed: the value 1 was observed at $0 \in \mathbb{R}^m$, 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 0 is the same as the impact of the same observation at 0 on x.

A4 Ray Invariance: Let there be given $B = (x_i, y_i)_{i \leq n} \in \mathbb{B}$, such that, for some $v \in \mathbb{R}^m$ and $\alpha_i \geq 0$ $(i \leq n)$, $x_i = \alpha_i v$. Then, for every $d > 0$, $Y((x_i + dv, y_i)_{i \leq n}, 0) = Y((x_i, y_i)_{i \leq n}, 0)$.

A4 applies only in the very special case in which all the points in the given database are ordered on a ray emanating from the origin, $\{\, x + \lambda v \,|\, \lambda \geq 0 \,\}$. In this case, A5 requires that moving all the points in the database further away along the ray (without changing the distances between them) will not change the prediction of Y.

A5 Self-Relevance: For every $x, z \in \mathbb{R}^m$, $Y(((0, 1), (x, 0)), z) \leq Y(((0, 1), (x, 0)), 0)$.

A5 considers a simple database B consisting of two points: the value 1 was observed for the point 0, while the value 0 was observed for the point x. Given such a database, any prediction generated by equation (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.

A *norm* is a function $n : \mathbb{R}^m \to \mathbb{R}_+$ satisfying:
(i) $n(\xi) = 0$ iff $\xi = 0$;
(ii) $n(\lambda\xi) = |\lambda| n(\xi)$ for all $\xi \in \mathbb{R}^m$ and $\lambda \in \mathbb{R}$;
(iii) $n(\xi + \zeta) \leq n(\xi) + n(\zeta)$ 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 above. The following are equivalent:*

(i) Y satisfies A1-A5;

(ii) There exists a norm $n : \mathbb{R}^m \to \mathbb{R}_+$ such that

$(*)$ $\quad s(x, z) = \exp[-n(x - z)]$ $\quad$ *for every* $x, z \in \mathbb{R}^m$

We observe that, given s, the norm n 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, Lieberman, and Schmeidler (2004). 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, Lieberman, and Schmeidler has no structure whatsoever. The only property that follows from their axiomatization An important feature of our result is that observable conditions

on predictions Y imply that n is a norm, and this, in turn, imposes restrictions on the similarity function. First, since for a norm n, $n(\xi) = n(-\xi)$, 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 n implies that 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 $B = (x_i, y_i)_{i \leq n}$, $Y((x_i, y_i)_{i \leq n}, 0) = Y((x_i P, y_i)_{i \leq n}, 0)$.

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 n 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. We 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 $y \in \mathbb{R}$, $Y((B, (x, y)), 0) = Y((B, (x', y)), 0)$, where $(B, (x, y))$ denotes the database obtained by concatenation of B with (x, y). In light of equation (1), it is easy to see that two vectors x and x' are considered

$\sim$-equivalent if and only if $s(x,0) = s(x',0)$. 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 $\theta, \mu > 0$ be such that $\beta\theta^2 + \mu^2 = \beta$. Then for every $x = (x^1, ..., x^m)$, $x + e_j \sim x + \theta e_j + \mu e_k$.

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 n 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 Appendix: Proof

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.[2]

It is easy to see that A1 is equivalent to the existence of a function $f : \mathbb{R}^m \to \mathbb{R}_{++}$, with $f(0) = 1$, such that $s(x,z) = f(x - z)$ for every

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

$x, z \in \mathbb{R}^m$. Indeed, if such an f exists, A1 will hold. Conversely, if A1 holds, one may define $f(x) = s(x, 0)$ 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 $f(x) = s(x, 0)$, 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 $x \in \mathbb{R}^m$, and $f(0) = 1$.

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, $\{ \lambda x \mid \lambda \geq 0\}$, for a given $x \in \mathbb{R}^m$ ($x \neq 0$). We observe that for Ray Invariance to hold, in the presence of Monotonicity, $s(\lambda x, 0)$ has to be exponential in λ. To see this, observe that Ray Invariance implies that the ratio $s(k\lambda x, 0)/s((k+1)\lambda x, 0)$ is independent of k for every λ. This guarantees that $s(\lambda x, 0)$ 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 n_x such that $s(\lambda x, 0) = \exp[-\lambda n_x]$. Obviously, $n_{\lambda x} = \lambda n_x$ for $\lambda \geq 0$. A2 also implies that $n_x > 0$ for $x \neq 0$.

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 $s(x, z) = f(x - z)$ every $x, z \in \mathbb{R}^m$, where $f(0) = 1$, $f(x) = f(-x)$ for every $x \in \mathbb{R}^m$, and, for every $x \in \mathbb{R}^m$ there exists a non-negative number n_x such that $f(x) = \exp[-\lambda n_x]$ and $n_{\lambda x} = \lambda n_x$ for $\lambda \geq 0$. Further, $n_x = 0$ only for $x = 0$. Defining $n(x) = n_x$ we obtain the representation $(*)$ for a function n that satisfies all the condition of a norm, apart from the triangle inequality.

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

$$Y(((0, 1), (x, 0)), z) \leq Y(((0, 1), (x, 0)), 0)$$

which implies that

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

or

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

Equivalently, we have

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

which is equivalent, in turn to

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

and to

$$s(x,z) \geq s(x,0)s(0,z)$$

Observe that A5 is equivalent to this form of multiplicative transitivity independently of the other axiom. While we obtain the multiplicative transitivity condition only at 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 $x,z \in \mathbb{R}^m$,

$$\exp[-n(x-z)] \geq \exp[-n(x)-n(-z)]$$

or

$$n(x-z) \leq n(x)+n(-z)$$

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

This completes the proof of the theorem.□

References

Akaike, H. (1954), "An Approximation to the Density Function", *Annals of the Institute of Statistical Mathematics*, **6**: 127-132.

Billot, A., I. Gilboa, D. Samet, and D. Schmeidler (2003), "Probabilities: Frequencies viewed in Perspective", mimeo.

Gilboa, I. and D. Schmeidler (2003) "Inductive Inference: An Axiomatic Approach", *Econometrica*, 71 (2003), 1-26.

Gilboa, I., O. Lieberman, and D. Schmeidler (2004) "Empirical Similarity", mimeo.

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*. New York: John Wiley and Sons.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*. London and New York: Chapman and Hall.