EMPIRICAL SIMILARITY

Itzhak Gilboa, Offer Lieberman, and David Schmeidler*

Abstract—An agent is asked to assess a real-valued variable Y_p based on certain characteristics $X_p = (X_p^1, \ldots, X_p^m)$, and on a database consisting of $(X_i^1, \ldots, X_i^m, Y_i)$ for $i = 1, \ldots, n$. 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, $\bar{Y}_p^s$, be the weighted average of all previously observed values Y_i, where the weight of Y_i for every $i = 1, \ldots, n$, is the similarity between the vector $X_p^1, \ldots, X_p^m$, associated with Y_p, and the previously observed vector, $X_i^1, \ldots, X_i^m$. We axiomatize this rule. We assume that, given every database, a predictor has a ranking over possible values, and we show that certain reasonable conditions on these rankings imply that they are determined by the proximity to a similarity-weighted average for a certain similarity function. The axiomatization does not suggest a particular similarity function, or even a particular form of this function. We therefore proceed to suggest that the similarity function be estimated from past observations. We develop tools of statistical inference for parametric estimation of the similarity function, for the case of a continuous as well as a discrete variable. Finally, we discuss the relationship of the proposed method to other methods of estimation and prediction.

I. Introduction

A. Motivation

ECONOMIC agents as well as various professionals are often required to assess the value of a certain numerical variable. In many situations, available data are relevant for the assessment problem, but they do not suggest a value that is indisputably the only reasonable assessment to make. Consider the following examples.

1. A homeowner considers selling her house, and she wonders how much she could get for it. Naturally, she should be basing her assessment on the prices at which other houses were sold. Yet, every house has its idiosyncratic characteristics. Hence the "market value" of her house is a variable that needs to be assessed based on observations of other transactions, but cannot be uniquely determined by these transactions in the same way that the price of a ton of wheat can.
2. An art dealer wants to sell a painting by a fairly famous painter. Evidently, the market price of the painting is related to the prices at which other, similar paintings were sold. Yet, the painting is unique, and its price may differ from the prices of all other paintings, as well as from their average.
3. An analyst is asked to predict the rate of inflation for the coming year. Using past empirical frequencies of various inflation rates is hardly an option in this case, because every year differs from past years in several ways. Yet, it is obvious that past inflation rates are informative and should somehow be used for the prediction.[1]
4. The same analyst is now asked to assess the probability of a stock market crash within the next six months. Again, she is expected to generate an assessment that is based on past observations. However, every two situations would typically differ in the values of certain important economic variables.
5. A physician is asked to assess the probability of success of an operation to be performed on a certain patient. Past experience with other patients is clearly relevant and should inform the assessment process. Yet, every human body is unique, and simple relative frequencies of success do not summarize all the relevant information.
6. A lawyer is asked by her client what are the chances of winning a case. Clearly, every case is idiosyncratic. Yet, the rulings in similar cases and under like-minded judges are relevant for the assessment.

In all of these problems one attempts to assess the value of a variable Y_p based on the values of relevant variables, $X_p = (X_p^1, \ldots, X_p^m)$, and on a database consisting of the variables $(X_i^1, \ldots, X_i^m, Y_i)$ for $i = 1, \ldots, n$. The question is, how do and how should people combine past observations of X and Y with the current values of X to generate an assessment of Y?

This problem is extensively studied in statistics, machine learning, and related fields. Among the numerous methods that have been suggested and used to solve such problems one may mention parametric and nonparametric regression, neural nets, linear and nonlinear classifiers, k-nearest-neighbor approaches (Fix & Hodges, 1951, 1952; Cover & Hart, 1967; Devroye, Gyorfi, & Lugosi, 1996), kernel-based estimation (Akaike, 1954; Rosenblatt, 1956; Parzen, 1962; Silverman, 1986; Scott, 1992), and others. Each of these methods has considerable success in a variety of applications. Moreover, each method can also be viewed as a tentative model of human reasoning. How should we choose among these approaches for descriptive and for normative applications?

Received for publication December 7, 2004. Revision accepted for publication September 19, 2005.

* Tel-Aviv University, HEC, and Yale University; The Technion; and Tel-Aviv University and The Ohio State University, respectively.

We wish to thank John Geanakoplos and Don Brown for conversations that greatly influenced this work. We are also grateful to Daron Acemoglu, Yoav Binyamini, Raul Drachman, Gabi Gayer, Mark Machina, Enno Mammen, Yishay Mansour, and two anonymous referees for comments and references. Gilboa and Lieberman gratefully acknowledge support from the Pinhas Sapir Center for Development and the Chaim Katzman Gazit-Globe Real Estate Institute at Tel-Aviv University. 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 the Israel Science Foundation Grant 975/03.

[1] This application was suggested by Raul Drachman.

The Review of Economics and Statistics, August 2006, 88(3): 433–444

Our approach to this problem is axiomatic and empirical. We start with a system of axioms that characterizes a class of assessment rules. We do not expect the axiomatic approach or other theoretical considerations to fully specify the parameters of the assessment rule. Rather, we suggest that these parameters be estimated from data. This estimation is done in the context of a probability model that allows statistical inference. We now turn to describe this approach in more detail.

B. Axiomatization of Similarity-Weighted Averaging

For the axiomatic model we assume that, given a database $B = (X_i, Y_i)_{i \leq n}$, where $n \in \mathbb{N}$, $X_i \in \mathbb{R}^m$, and $Y_i \in \mathbb{R}$, and a new data point $X_p \in \mathbb{R}^m$, the agent has a ranking $\gtrsim_{B,X_p}$ over the possible values of Y_p. The interpretation of $\xi \gtrsim_{B,X_p} \zeta$ is that, given the database B and the new data point X_p, ξ is more likely to be observed than is ζ. We study the rankings $\gtrsim_{B,X_p}$ that the agent would generate given various possible databases, holding m fixed. We formulate axioms on such rankings, and show that the rankings satisfy these axioms if and only if they can be represented by similarity-weighted averaging. Specifically, the axioms are equivalent to the existence of a function $s : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_{++} = (0, \infty)$ such that, given a database $B = (X_i, Y_i)_{i \leq n}$ and a new data point $X_p = (X_p^1, \ldots, X_p^m) \in \mathbb{R}^m$, two possible estimates of Y_p are ranked according to their proximity to the similarity-weighted average of all observations in the database, namely,

$$\bar{Y}_p^s = \frac{\sum_{i \leq n} s(X_i, X_p) Y_i}{\sum_{i \leq n} s(X_i, X_p)}. \tag{1}$$

This rule for generating predictions is reminiscent of kernel estimation. [See Akaike (1954), Rosenblatt (1956), and Parzen (1962). See details in Section II below.] We prefer the term "similarity" because it suggests a cognitive interpretation of the function, as opposed to the more technical "kernel." This is obviously only a matter of interpretation.[2]

The axioms we propose are not universal, and they need not be satisfied by all types of human reasoning. Specifically, when people use the data to develop theories, and then use these theories to generate predictions, they are unlikely to satisfy our axioms, or to follow equation (1). (We elaborate on this point after the presentation of the axioms in section III.) Our axioms attempt to describe the assessment of an agent who aggregates data, but who does not engage in theorizing. When agents do reason by general rules, or theories, a model such as regression analysis may be a better model than the similarity-weighted averaging we discuss here.

[2] Our axiomatization relies on that of Gilboa and Schmeidler (2001, 2003). Yet, the former is not a special case of the latter. Moreover, the analysis conducted here employs the fact that the variable Y is real-valued.

We also axiomatize the relation "more likely than" that corresponds to a set of agents, constituting a "market," and we show that, under our axioms, one may replace all agents with their subjective similarity functions by a "representative" agent with an appropriately defined similarity function.

C. The Empirical Similarity

The axiomatization we propose does not specify a particular similarity function, or even a particular functional form thereof.[3] Where do the similarity numbers come from? In this paper we do not attempt to provide a theoretical answer to this question. Rather, we suggest an empirical approach: given a database $B = (X_i, Y_i)_{i \leq n}$, we assume that past values Y_i were also generated in accordance with equation (1), adapted for $p = i$ and $n = i - 1$, that is,

$$\bar{Y}_i^s = \frac{\sum_{k<i} s(X_k, X_i) Y_k}{\sum_{k<i} s(X_k, X_i)}, \tag{2}$$

relative to the similarity function s of the representative agent. We then ask which similarity function $s : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_{++}$ can best fit the data B under this assumption. This function, dubbed *the empirical similarity,* can then be used to generate assessments of $\bar{Y}_p^s$. These assessment will be more objective than similar assessments based on a subjective similarity function.

In this paper we address a parametric version of the question of estimation of the similarity function. We suggest a functional form of s, and estimate its parameters with a maximum likelihood estimator in a statistical model that we define shortly. However, an "empirical similarity function" may be any function that is estimated from the data, or that is chosen to fit the data according to equation (2).

Further discussion of our estimation method and the assumptions underlying it is deferred to section VI. We now proceed to describe a statistical model within which this estimation can be analyzed.

D. Statistical Analysis

The empirical similarity we obtain can be viewed as a point estimate of a similarity function, if we embed equation (1) in a statistical model. Specifically, we are interested in similarity functions that depend on a weighted Euclidean distance,

$$d_w(x, x') = \sqrt{\sum_{j \leq m} w_j (x^j - x'^j)^2}, \tag{3}$$

[3] Billot, Gilboa, and Schmeidler (2004) offer an axiomatization of a particular functional form of a similarity function. Assuming that an agent employs a similarity-weighted averaging as suggested here, they impose additional axioms on the agent's assessments given various databases and various new data points, which are equivalent to the existence of a norm on R^m such that the similarity function is a negative exponential in this norm.

where $x = (x^1, \ldots, x^m)$ and $x' = (x'^1, \ldots, x'^m)$. The similarity function may be expected to decrease in the distance d_w, to attain the value 1 for $d_w = 0$, and to converge to 0 as $d_w \to \infty$. Natural candidates for such a function include $s_w = e^{-d_w}$ and $s_w = 1/(1+d_w)$. Billot et al. (2004) assume that an agent generates assessments according to equation (1), and take an axiomatic approach to the problem of selecting the functional form of the similarity function. Specifically, they show that certain conditions on the assessments Y_p generated given various databases are equivalent to the existence of a norm $\|.\|$ on $\mathbb{R}^m$ such that $s_w(x,x') = e^{-\|x-x'\|}$. Because d_w is a norm on R^m when $w_j > 0$ for all $j \leq m$, the function $s_w = e^{-d_w}$ may be viewed as a special case of the similarity function axiomatized in Billot et al. (2004).

Observe that the weights $(w_j)_j$ are not restricted to sum to 1. This allows some flexibility in the relative weight of closer versus more remote observations. For instance, multiplying all weights $(w_j)_j$ by a constant $\lambda^2 > 0$ is tantamount to multiplying d_w by $\lambda > 0$. If $\lambda > 1$, this transformation reduces the relative effect of remote points.

For $t = 2, \ldots, n$, we assume that

$$Y_t = Y_t^s = \frac{\sum_{i<t} s(X_i, X_t) Y_i}{\sum_{i<t} s(X_i, X_t)} + \varepsilon_t, \tag{4}$$

where $\varepsilon_t \sim N(0, \sigma^2)$, and Y_1 is an arbitrary random variable.

In such a model it makes sense to ask whether the point estimates of the unknown parameters are significantly different from a prespecified value, and in particular, from 0. In this paper we focus on maximum likelihood estimation of the parameters $(w_j)_j$, and we develop tests for such hypotheses.

For some applications, including examples 5 and 6 above, the observed values of Y_t are categorical. In this case one cannot assume a model such as equation (4), and the latter should be replaced with a model of the form

$$P(Y_t = 1 | X_1, Y_1, \ldots, X_{t-1}, Y_{t-1}) = F(Y_t^s),$$

where F is a cumulative distribution function, X_i is an m-vector, and $Y_i \in \{0, 1\}$, with $Y_i = 1$ denoting success and $Y_i = 0$ denoting failure in examples 5 and 6. This model differs from discrete choice models in a way that parallels the difference between our model for a continuous Y_t^s and linear regression. Specifically, the probability that Y_t assumes the value 1 depends on the weighted relative frequency of 1 among past values $\{Y_i\}_{i<t}$, where the weight of the value Y_i depends on the similarity between the vector X_i observed in the past and the current observation X_t. We provide a statistical model for this case, and develop tests for hypotheses about the values of the parameters $(w_j)_j$ in this model as well.

The rest of this paper is organized as follows. In section II we discuss the relationship between our method and existing statistical methods. Section III provides the axiomatization of similarity-weighted averaging, for a single agent and for a set of agents. In section IV we develop the statistical theory for the continuous case, whereas section V deals with the discrete case. Finally, section VI concludes.

II. Related Techniques

Our main focus is on human reasoning. We are interested in data that are generated by people, and we take the similarity-weighted average as a possible model of how people generate assessments. That is, we interpret our model as describing a causal relationship.

Our method can be applied also to databases in which the variable Y is not a result of human reasoning. In this case our model should not be interpreted causally, but one may still find a similarity function that best fits the data. Moreover, one may even conduct hypothesis tests for the parameters of the similarity function, to the extent that one believes that the data-generating process may be in agreement with one of the models specified above. In other words, the empirical approach suggested here, coupled with the statistical inference that accompanies it, may be viewed as a general-purpose statistical technique dealing with the prediction of a variable Y based on variables $X_1, \ldots, X_m$ and past observations of all these variables in conjunction.

Viewed from this perspective, one might wonder how our prediction technique compares with established ones, such as regression analysis. An obvious weakness of our approach is that it does not attempt to identify trends. For instance, assume that there exists a single variable X which denotes time, and that the data lie on a line $Y = X$. This obvious trend will not be recognized by our technique, which will continue to expect the next value of Y to be a weighted average of past values of Y. The prediction technique we suggest makes sense especially when one might believe that past observations were obtained under similar circumstances.

A. *Nonlinear Regression*

Our approach differs from nonlinear regression in that we do not assume that the data-generating process follows a basic functional relationship of the form $Y = f(X^1, \ldots, X^m)$. Rather, we assume that Y is distributed around a weighted average of its past values, where the X's determine these weights.

If, however, one does assume that there exists an underlying functional relationship $Y = f(X^1, \ldots, X^m)$, our technique may still be used for prediction of Y. As long as f is sufficiently smooth, one may hope that, with a large number of observations that are evenly scattered with respect to their X values, the similarity-weighted averaging will result in reasonable predictions. Indeed, the similarity-weighted average is reminiscent of the Nadaraya-Watson estimator of a nonparametric functional relationship. Observe that, as opposed to the Nadaraya-Watson technique and related literature, we do not attempt to find an optimal kernel function based on theoretical considerations, but find the kernel (similarity function) that best fits the data.

Our estimation of the similarity function s is parametric. This does not imply that we restrict the function f to a parametrized family of functions, should a relationship $Y = f(X^1, \ldots, X^m)$ actually exist. Whereas any function s may be used to generate predictions in a nonparametric problem, we restrict attention to functions s within a parametrized family of similarity functions. Thus, we try to parametrically estimate how to best perform nonparametric estimation.

B. *Kernel Estimation and Case-Based Reasoning*

If we think of similarity-weighted averaging as a model of human reasoning, we find that a case-based reasoner, as modeled by this formula, can be viewed as someone who believes in a general rule of the form $Y = f(X^1, \ldots, X^m)$ but does not know the functional form of f and therefore attempts to estimate it by nonparametric techniques.

The notion that people reason by analogies dates back to Hume (1748) at the latest. In artificial intelligence, this idea was revived as case-based reasoning by Schank (1986) and Riesbeck and Schank (1989). Inspired by this work, Gilboa and Schmeidler (1995, 2001, 2003) developed a formal, axiomatically based theory of decision and prediction by analogies. In this literature it has been mentioned that case-based reasoning is a natural and flexible mode of thinking and decision-making. Our statistical approach strengthens this intuition by pointing out that case-based reasoning may be a way to estimate a functional rule.

Taking an evolutionary viewpoint, assume that nature programs the mind of an organism who needs to operate in an unknown environment. The organism will need to learn certain functional rules of the form $Y = f(X^1, \ldots, X^m)$, but it is not yet known what form the function f might take. The statistical viewpoint suggests case-based assessment by similarity-weighted averaging as a procedure to predict Y, which may perform well in a variety of possible environments f. Moreover, it turns out that the similarity-weighted averaging does not explicitly resort to general rules and theories, and thus does not require abstract thinking. Case-based reasoning therefore appears to be a flexible methodology of learning rules, which can be implemented on simple machines. Admittedly, this method is limited, and human reasoning requires also abstract thinking and the development of explicit general theories. Yet, the evolutionary viewpoint seems to support case-based reasoning as a simple but powerful technique.

C. *Interpolation*

Our prediction method can also be viewed as a type of interpolation. Consider first the case $m = 1$, that is, a single variable X. Every past case is a point $(x_i, y_i) \in \mathbb{R}^2$, and we are asked to assess the value of Y for a new point $x_p \in \mathbb{R}$. Assume for simplicity that x_p is in the interval $[\min_i x_i, \max_i x_i]$. Linear interpolation would generate a prediction by the line segment connecting (x_i, y_i) and (x_k, y_k) for the two values x_i and x_k that are closest to x_p in either direction. This approach may be a bit extreme in that it uses only the y values for the closest x's. In this respect, it is similar to a (single) nearest-neighbor technique. Other types of interpolation, such as polynomial interpolation, would take into account also other points (x_l, y_l) for x_l that is not necessarily the closest to x_p on either side.

These interpolation techniques implicitly assume that the values observed are the actual, precise values of an unknown function. If, however, we recognize that the process contains some inherent randomness, that we may not measure certain hidden variables, or that there are measurement errors, we might opt for a technique that is less sensitive to each particular value of Y. Following this line of thought, our approach can be viewed as performing *statistical interpolation:* every observation is used in the interpolation process, where closer points have a larger effect on the predicted value. As opposed to interpolation by high-order polynomials, when many points have been observed, no particular point would have a large effect on the predicted value.

When we consider the case $m > 1$, generalizing this interpolation technique requires a multidimensional distance function. Our methodology might therefore be conceptualized as a multidimensional statistical interpolation technique, where the distance function is empirically learnt.

D. *Bayesian Updating*

A special case of the formula (1) is when s is constant (say, $s \equiv 1$), and the formula boils down to the simple average (of Y) over the entire database. This could be viewed as an estimator of the unconditional expectation of Y, having not observed any X's.

By contrast, one may consider an extreme similarity function given by $s(X_i, X_p) = 1_{\{x_i = x_p\}}$, where 1 denotes the indicator function. That is, two data points are considered to be perfectly similar if they have exactly the same X values, and absolutely dissimilar otherwise.[4] In this case, formula (1) yields the average of Y over the subdatabase defined by the values X_p, and it can be viewed as an estimate of the *conditional* expectation of *Y, given* X_p.

Thus, the formula (1) provides a continuum between conditional and unconditional expectations. When $s(X_i, X_p) = 1_{\{x_i = x_p\}}$, the reasoner only considers identical cases as relevant, and all of them are then deemed equally relevant. By contrast, if $s \equiv 1$, the reasoner considers all cases as identically relevant. In between, equation (1) allows for various cases to have various degrees of relevance. Given the new data point X_p, past points X_i are judged for their

[4] In our model the similarity function is positive everywhere. This simplifies the formula and the axiomatization alike. But one can extend the model to include similarity functions that may vanish, or consider zero similarity values as a limit case.

relevance, but not in a dichotomous way. In other words, Bayesian updating may be viewed as a special case of equation (1), where similarity is evaluated in a binary way: two observations are similar if and only if they are identical in every possible known aspect.

As compared to Bayesian updating, a reasoners who employs equation (1) might be viewed as a less extreme assessor of similarity. She does not use only the observations with identical X values, but also other, less relevant ones. Why would she do that? Why should she contaminate her assessment of Y for X_p with Y's that were observed for other X's?

The answer is, presumably, the scarcity of data. If we are faced with a database in which the very same X_p values appear a very large number of times, it would seem reasonable to assess the conditional expectation of Y given X_p based solely on the observations that share the exact values of X_p. But one may find that these exact values were encountered very few times, if at all. Indeed, the X's might include certain variables, such as time and location, that uniquely identify the observation. In this case, no two observations ever share the exact X values, and conditioning on X_p leaves one with an empty subdatabase. Even in less extreme examples, the resulting subdatabase may be too meager for generating predictions. In those cases, the formula (1) offers an alternative, in which the similarity of the observations is traded off for the size of the database.

Viewed thus, formula (1) may deserve the name "kernel updating." As in other kernel-based techniques, the relevance of an observation (X_i, Y_i) is not restricted to identical data points $X_p = X_i$, but is extended to other data points X_p, to an extent determined by the kernel values $s(X_i, X_p)$. The use of a kernel function in this case is justified by the paucity of the data, that is, by the fact that observations with precisely the same X_p are scarce. This parallels the motivation for the use of kernel functions in kernel estimation of a density function and in kernel classification.

Finally, we observe that the use of observations (X_i, Y_i) where $X_i \neq X_p$ for the prediction of Y_p may also follow from Bayesian updating if one assumes that the X variables are observed with noise.[5]

E. Autoregression Models

From a mathematical viewpoint, the similarity-weighted average can be regarded as a type of an autoregression model. In autoregression models, as well as in our case, Y_t is distributed around a linear function of past values of Y.[6] Yet, the similarity-weighted average formula differs from autoregression models in several important ways. Mathematically, the weights that past values $\{Y_i\}_{i<t}$ have in the equation of Y_t do not depend on the time difference $(t - i)$, but on the similarity of the corresponding X values, that is, on $s(X_i, X_t)$. In particular, observe that the weights of $\{Y_i\}_{i<t}$ in the determination of the expectation of Y_t are not known before time t, because these weights depend on X_t. Observe also that in our case each Y_t depends on *all* past observations. Thus, our model is an autoregression model whose order is not bounded a priori. Another important difference is that in our case the index t has no cardinal significance. We use it only to order the data; our procedure does not rely on the fact that the time difference between observations $t - 1$ and t is the same as the time difference between observations $t - 2$ and $t - 1$.[7]

Conceptually, our model assumes that similar situations in the past might have a significant effect on current values of Y, even if they occurred a long time ago. When one discusses natural phenomena, such as population growth, one expects the weight of past observations to be increasing as a function of their recency. But when we deal with human reasoning, as in the case of inflationary expectations, less recent, but more similar situations in the past may have a greater influence on the future than would more recent but less similar situations. In a sense, human memory may serve as a channel through which past periods can affect future periods without the mediation of the periods in between.

The above need not imply that our model ignores time completely. One may introduce time as one of the variables X^j. This would allow more recent periods to have greater influence on the prediction than less recent ones, simply because the time difference is translated, via the variable X^j, to a distance in the X space, and thus to a lower degree of similarity.

F. How to Analyze Time Series

We conclude that the resemblance between our model and autoregression models is superficial. Yet, our model can be adapted to deal with time series in a way that resembles auto regression in a more profound way. Autoregression can be viewed, in bold strokes, as explaining a variable by its own past values, with statistical techniques such as linear regression. The natural counterpart in our case would be to predict the variable Y by equation (1) where the variables $(X^j)_j$ include lagged values of Y itself. For example, assume that Y_t is a quarterly growth rate. Introducing $Y_{t-1}, \ldots, Y_{t-k}$ as $X_t^1, \ldots, X_t^k$ would suggest that the predicted rate of growth at period t be a (weighted) average of the rates of growth in similar periods in the past, *where similarity is defined by the pattern of growth rates in the most recent k periods.* Our technique would find weights $w_1, \ldots, w_k$ that best fit the data when one uses the equation

[5] This comment is due to Mark Machina.

[6] As pointed out to us by an anonymous referee, when the similarity function is allowed to vanish, the i.i.d. process is a special case of our process when $s(X_1, X_t) = 1$ and $s(X_j, X_t) = 0$ for $1 < j < t$, and $Y_1 = 0$.

[7] In fact, our procedure can be easily adapted to the case in which observations are only partially ordered. As we briefly mention below, a different variant of our model can deal with situations in which the observations are not ordered at all.

$$\bar{Y}_t^s = \frac{\sum_{i<t} s_w((Y_{i-k}, \ldots, Y_{i-1}), (Y_{t-k}, \ldots, Y_{t-1}))Y_i}{\sum_{i<t} s_w((Y_{i-k}, \ldots, Y_{i-1}), (Y_{t-k}, \ldots, Y_{t-1}))}, \quad (5)$$

where

$$s_w((Y_{i-k}, \ldots, Y_{i-1}),(Y_{t-k}, \ldots, Y_{t-1})) = \exp\Big(-\sqrt{\sum_{j\leq k} w_j(Y_{i-j} - Y_{t-j})^2}\Big).$$

This estimation technique could be interpreted as follows. We first ask, what determines the similarity of patterns of growth? That is, is a "pattern" defined by the most recent period or, if by several most recent periods, how many of these, and what are their relative weights? The estimation of the weights w_j attempts to answer this question. Although the resulting weights need not be monotonically decreasing in j (the time difference), one would expect that these weights would become small for large values of j. In fact, in determining the number of periods that define a "pattern," k, one implicitly assumes that periods more distant than k are not part of the "pattern." The selection of this k may be compared to the selection of the order p in autoregression models of order p [AR(p)].

Once the weights w_j have been determined, we search the entire database for periods i such that the pattern preceding i, $(Y_{i-k}, \ldots, Y_{i-1})$, resembles the current pattern, $(Y_{t-k}, \ldots, Y_{t-1})$. For such periods, the value Y_i would have a larier-weight in the prediction of Y_t than would the value corresponding to periods for which $(Y_{t-k}, \ldots, Y_{t-1})$ resembles $(Y_{t-k}, \ldots, Y_{t-1})$ to a lesser degree. Again, one may also add time as an additional variable X^{k+1} to make sure that the prediction discounts the past.

III. Axiomatization

A. Single Agent

The axiomatization does not require that past data points range over all of $\mathbb{R}^m$. We assume that they belong to a nonempty subset $\Gamma \subset \mathbb{R}^m$. However, we do assume that every possible data point in Γ may have been observed together with every value $y \in \mathbb{R}$ any finite number of times. We therefore model the database as a vector of counters, denoted I, rather than the set of observations B used in the introduction.

Specifically, let $C = \Gamma \times \mathbb{R}$ denote *case types.* A case type $(x, \xi) \in C$ is interpreted as an observation of a *data point* $x \in \Gamma$ coupled with the value $\xi \in \mathbb{R}$. *Memory* is a nonzero function $I : C \rightarrow \mathbb{Z}_+$ (where $\mathbb{Z}_+$ denotes the nonnegative integers) such that $\sum_{c\in C} I(c) < \infty$, specifying for every case type c how many cases of that type have appeared. Let $\mathcal{I}$ be the set of all memories.

We are currently presented with a new data point $x_p \in \Gamma$. The task is to estimate the value $\eta \in \mathbb{R}$ that corresponds to x_p. We assume that the predictor not only chooses one such η, but has a likelihood ranking over all possible predictions. Formally, for $I \in \mathcal{I}$, let $\gtrsim_1 \subset \mathbb{R} \times \mathbb{R}$ be a binary relation over the reals. As usual, $>_I$ denotes the asymmetric part of $\gtrsim_1$. For $\xi,\eta \in \mathbb{R}$, $\xi \gtrsim_1 \eta$ is interpreted as "Given the memory I, ξ is at least as likely as a value for the variable Y at the new data point x_p than is η." Observe that in the formal notation we suppress x_p. This new data point is fixed throughout this section.

We now state axioms on $\{\gtrsim\}_{I\in\mathcal{I}}$. The first three are identical to those appearing in Gilboa and Schmeidler (2001, 2003).

A1 Order: For every $I \in \mathcal{I}$, $\gtrsim_I$ is complete and transitive on $\mathbb{R}$.

A2 Combination: For every $I, J \in \mathcal{I}$ and every $\xi, \eta \in \mathbb{R}$, if $\xi \gtrsim_I \eta$ ($\xi >_I \eta$) and $\xi \gtrsim_J \eta$, then $\xi \gtrsim_{I+J} \eta$ ($\xi >_{I+J} \eta$).

A3 Archimedean Axiom: For every $I, J \in \mathcal{I}$ and every $\xi, \eta \in \mathbb{R}$, if $\xi >_I \eta$, then there exists $l \in \mathbb{N}$ such that $\xi >_{lI+J} \eta$.

Observe that in the presence of axiom A2, axiom A3 also implies that for every $I, J \in \mathcal{I}$ and every $\xi, \eta \in \mathbb{R}$, if $\xi >_I \eta$, then there exists $l \in \mathbb{N}$ such that for all $k \geq l$ we have $\xi >_{kI+J} \eta$.

Axiom A1 is rather standard. It requires that, given any memory, the "at least as likely as" relation be a weak order.

Axiom A2 is the main axiom of Gilboa and Schmeidler (1997, 2001, 2003). Roughly, it states that, if ξ is more likely than η given each of two memories, then ξ should also be more likely than η given their union. This axiom is satisfied by a variety of statistical techniques, such as kernel estimation, kernel classification, and maximum likelihood rankings. Yet, it is by no means universal. To illustrate its limitations, consider the following example. Suppose that there is only one predictor ($m = 1$), that the database consists of $\{(1, 1), (2, 2), \ldots, (5, 5)\}$, and the new data point is $X_6 = 6$. Given each observation (i, i) for $i = 1, \ldots, 5$, the value 6 might seem less likely than the value 5. But given the entire database, where five points lie exactly on the line $Y = X$, the value 6 seems a much more reasonable prediction for $X_6 = 6$. Indeed, the similarity-weighted average formula that we axiomatize is doomed to predict some weighted average of the values $\{1, \ldots, 5\}$, and will not be able to predict a value higher than 5.

This example shows a major limitation of the similarity-weighted formula, for which axiom A2 carries most of the blame: this formula is incapable of identifying trends and generating predictions based on them. Axiom A2 suggests that a conclusion that holds in two databases has to hold in their union. But if there is a trend, or a pattern in the data, it may be identified only when data are amassed. A2 rules out the possibility that the union of two memories would generate new insights. Similarly, if the similarity function is being learned by the predictor while she produces predictions, or if the estimator uses both inductive and deductive reasoning, then the combination axiom should not be expected to hold. Moreover, if the predictor knows that the data are generated by a particular model, such as a linear regression model or a specific Bayesian model, she will generate predictions based on that model. In this case she is likely to satisfy the combination axiom when estimating the parameters of the model (and, in particular, maximum likelihood estimation will satisfy the axiom), but not at the level of specific predictions generated by the model. In spite of all these limitations, A2 appears reasonable as a requirement on simple aggregation of evidence, in the absence of a theory on the way the data are generated.

A3 states that, if memory I contains evidence that ξ is more likely than η, then for each other memory J there exists a large enough number l such that l repetitions of I would be sufficient to overwhelm the evidence provided by J, and suggest that ξ is more likely than η also given the union of J and l times I. Thus A3 precludes the possibility that one piece of evidence is infinitely more weighty than another.

Gilboa and Schmeidler (1997, 2001, 2003) also use a diversity axiom, which we do not use here. Instead, we impose a new axiom that is specific to our setup. It states that, if memory I consists solely of cases that relate to the same data point x, then the ranking $\gtrsim_I$ is consistent with simple averaging. Observe that for such databases nothing can be learnt from the values of x, because they do not change at all. In this case, it makes sense that the most likely value of y be taken as the average of its observed values, and that possible values be ranked according to their proximity to this average.[8]

For $x \in \Gamma$, define $\mathcal{I}_x$ to be the set of memories in which only the data point x has been observed. Formally, $\mathcal{I}_x = \{I \in \mathcal{I} | I((x', y)) = 0 \text{ for } x' \neq x\}$. For $I \in \mathcal{I}_x$, define the average $y_I \in \mathbb{R}$ by

$$y_I = \frac{\sum_{(x,y) \in C} I((x,y))\, y}{\sum_{(x,y) \in C} I((x,y))}.$$

The last axiom we employ is

A4 Averaging: For every $x \in \Gamma$, every $I \in \mathcal{I}_x$, and every $\xi, \eta \in \mathbb{R}$, $\xi \gtrsim_I \eta$ iff $|\xi - y_I| \leq |\eta - y_I|$.
Our result can now be stated:

Theorem 1. Let there be given Γ, and $\{\gtrsim_I\}_{I \in \mathcal{I}}$. Then the following two statements are equivalent:

(i) $\{\gtrsim_I\}_{I \in \mathcal{I}}$ satisfy A1–A4,

(ii) There is a function s: $\Gamma \to \mathbb{R}_{++}$ such that

$$\text{for every } I \in \mathcal{I} \text{ and every } \xi, \eta \in \mathbb{R}, \quad \xi \gtrsim_I \eta \quad \text{iff} \quad |\xi - y_{s,I}| \leq |\eta - y_{s,I}|, \qquad (*)$$

where $y_{s,I} = \sum_{(x,y) \in C} s(x) I((x,y))\, y / \sum_{(x,y) \in C} s(x) I((x,y))$.

Furthermore, in this case the function s is unique up to multiplication by a positive number.

B. *Discussion*

The theorem states that, if we rank possible predictions of Y by their proximity to the average of past values of Y whenever the values of $X^1, \ldots, X^m$ are fixed, and we wish to extend this ranking to general databases in a way that satisfies our axioms (notably, the combination axiom), we are bound to do it by proximity to weighted averages.

The axiomatization we provide can be interpreted descriptively or normatively. From a descriptive point of view, the theorem suggests that, if an agent's rankings of possible values of a variable y given various databases satisfy our axioms, she can be ascribed a similarity function s such that her rankings are determined by proximity to a similarity-weighted average of past values of y, calculated by the similarity function s. From a normative viewpoint, the axiomatization might be used to convince an agent that similarity-weighted averaging is a reasonable way to assess the variable y given a database of past observations. Finally, the axiomatization also suggests a definition of an agent's similarity function, and method of elicitation for it.

A weighted averaging formula is also axiomatized in Billot et al. (2005). In their model a reasoner is asked to name a probability vector based on a memory I. Billot et al. impose an appropriate version of the combination axiom to conclude that the probability vector given a memory I is a weighted average of the vectors induced by each case in I separately. Unfortunately, the result of Billot et al. only applies if there are at least three states of the world, that is, if the probability vector has at least two degrees of freedom. For the special case of a single-dimension probability simplex, their theorem does not hold. In this respect, the present paper complements Billot et al. (2005).

C. *Representative Agent*

The theorem above shows under what conditions an agent's "as least as likely as" relation will follow the similarity-weighted average formula for an appropriately chosen similarity function. It relates the theoretical concept of similarity to the relation "at least as likely as," which is assumed to be observable.

In practice, however, one can often observe only aggregate data. For instance, one may observe market prices of houses or paintings, but not the assessments of these prices by agents. What properties should such assessments satisfy? How are individual assessments aggregated over agents?

[8] As pointed out to us by an anonymous referee, one may obtain axiomatizations of similarity-weighted versions of other statistics, such as the median. Any statistic that, in the absence of predictors, minimizes a convex cost function (summed over the given observations) may be viewed as the most likely value according to a relation $\gtrsim_I$ that satisfies A1–A3, and can then be generalized to a statistic that minimizes the similarity-weighted sum of that cost function.

Can such aggregates also be described as similarity-weighted averages?

To answer these questions, we extend the model presented above to incorporate more than one agent. Specifically, let $P = \{1, \ldots, p\}$ be a set of agents, and redefine the case types to be $C = P \times \Gamma \times \mathbb{R}$. A case of type (i, x, y) is interpreted as "agent $i \in P$ has observed a data point $x \in \Gamma$ and a corresponding value of $y \in \mathbb{R}$." Thus, every observation in this model specifies the observer, and not only the observed.

We continue as before to defined *memory* as a nonzero vector $I : C \to \mathbb{Z}_+$ such that $\sum_{c \in C} I(c) < \infty$. Let $\mathcal{I}$ be the set of all memories. We now think of a memory I is a matrix of counters, specifying how many times each agent has observed any possible $(x,y) \in \Gamma \times \mathbb{R}$ combination.

The relation $\gtrsim_I$ is interpreted as follows. For $\xi, \eta \in R$, $\xi \gtrsim_I \eta$ means that, if I specifies how many times each agent has seen each pair (x, y), then ξ is at least as likely as η to be the assessment of the set of agents. This assessment is supposed to reflect some collective opinion, and it does not reflect economic power or strategic considerations. If, for instance, we discuss the value of a painting by van Gogh, every agent is expected to have some assessment of the value of the painting, regardless of their ability or willingness to pay for it.

The axioms we use are the same axioms verbatim. The logic behind the axioms mirrors that of the single-agent case, though, naturally, in the multi agent case the axioms are more demanding.

We first state the theorem as applied to this case:

Corollary 2. Let there be given P, Γ, and $\{\gtrsim_I\}_{I \in \mathcal{I}}$. Then the following two statements are equivalent:

(i) $\{\gtrsim_I\}_{I \in \mathcal{I}}$ satisfy A1–A4.
(ii) There exist functions $\{s_i : \Gamma \to \mathbb{R}_{++}\}_{i \in P}$ such that

$$\text{for every } I \in \mathcal{I} \text{ and every } \xi, \eta \in R, \quad \xi \gtrsim_I \eta \quad \text{iff} \quad |\xi - y_{s,I}| \leq |\eta - y_{s,I}|, \tag{**}$$

where $y_{s,I} = \sum_{(i,x,y) \in C} s_i(x) I((i,x,y))\, y / \sum_{(i,x,y) \in C} s_i(x) I((i,x,y))$.

Furthermore, in this case the functions $\{s_i\}_{i \in P}$ are unique up to joint multiplication by a positive number.

We wish to show that, if we assume that all information is shared, then a set of agents P, characterized by functions $\{s_i\}_{i \in p}$, is indistinguishable from a representative agent whose similarity function is the average of $\{s_i\}_{i \in p}$. To this end, define $\mathcal{I}_{\text{sh}}$ as the set of memories in which all agents have the same information, that is: $\mathcal{I}_{\text{sh}} = \{I \in \mathcal{I} | I((i, x, y)) = I((i,x,y))$ for all $i, i' \in P\}$. We now have

Corollary 3. Let P, Γ, and $\{\gtrsim_I\}_{I \in \mathcal{I}}$ be given. Assume that $\{\gtrsim_I\}_{I \in \mathcal{I}}$ satisfy A1–A4. Then there exists a function s: $\Gamma \to \mathbb{R}_{++}$ such that

$$\text{for every } I \in \mathcal{I}_{sh} \text{ and every } \xi, \eta \in R, \quad \xi \gtrsim_I \eta \quad \text{iff} \quad |\xi - y_{s,I}| \leq |\eta - y_{s,I}|, \tag{***}$$

where $y_{s,I} = \sum_{(i,x,y) \in C} s(x) I((i,x,y))\, y / \sum_{(i,x,y) \in C} s(x) I((i,x,y))$. Furthermore, in this case the function s is unique up to multiplication by a positive number.

Observe that the identification of individual similarity functions s_i requires that memories $I \notin \mathcal{I}_{\text{sh}}$ be considered, that is, memories in which different agents may have observed different cases. Measuring $\gtrsim_I$ for $I \notin \mathcal{I}_{\text{sh}}$ and testing our axioms may be done in controlled experiments in a laboratory. It is more challenging to observe $\gtrsim_I$ for $I \notin \mathcal{I}_{\text{sh}}$ in empirical data. Yet, one may imagine that such relations exist and satisfy our axioms.

As long as we restrict attention to shared information, namely, to memories in $\mathcal{I}_{\text{sh}}$, we only observe the average similarity function. Attributing this average similarity to a representative agent, we conclude that the assessment made by a set of agents will be equivalent to that made by the representative agent.

The observability of $\gtrsim_I$ mirrors the observability of a utility function in economics: in principle, one may measure each agent's utility function. In reality, often only aggregate data are available. Under certain conditions, one may assume that the decisions of a set of agents can be described by the decision of a single, representative agent. Similarly, in our case one may, in principle, measure each agent's similarity function. In practice, we often observe only aggregate assessments. However, under the conditions specified above, we may replace the set of agents by a single, representative agent, and obtain the same assessment for shared information. It is this similarity function, of a representative agent, that we attempt to estimate.

IV. Statistical Inference for a Continuous Model

A. *The Model and the Likelihood Function*

If we assume the initial condition to be $Y_1 = \varepsilon_1$, then equation (4) can be written in matrix form as

$$Sy = \varepsilon,$$

where $S = S(w) = I - B_w A_w$, I is the identity matrix of order n;

$$A_w = \begin{pmatrix} 0 & & & \\ s_{w,2,1} & 0 & & \\ \vdots & & \ddots & \\ s_{w,n,1} & \cdots & s_{w,n,n-1} & 0 \end{pmatrix};$$

$$B_w = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & (e_2' A_w 1)^{-1} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & (e'_n A_w 1)^{-1} \end{pmatrix};$$

$S_{w,i,j} = s_w(X_j, X_i) = e^{-d_w(x_j, x_i)}$, d is based on equation (3): 1 is an $n \times 1$ vector of 1's; e_j is the canonical vector of 0's, apart from the jth position, where it is set to unity; $y = (Y_1, \ldots, Y_n)'$; and ε is an $n \times 1$ vector of i.i.d. Gaussian variables with mean 0 and variance σ^2. Note that S is a lower triangular matrix that does not depend on the variables Y_i.

We set $\theta = (\theta_1, \ldots, \theta_{m+1}) = (\sigma^2, w_1, \ldots, w_m)$ and observe that $\theta \in \Theta \subset \mathbb{R}_+^{m+1}$. The maximum likelihood estimator (MLE) of θ, $\theta_{\hat{n}}$, maximizes

$$l(\theta) = -\frac{n}{2}\log(2\pi\theta_1) - \frac{1}{2}y'H(\theta)y,$$

where $H = S'S/\sigma^2$.

Note the difference between nonparametric regression and our approach. In the former, the postulated relationships are of the form $y = g(x) + \varepsilon$ and the Nadaraya-Watson estimator of the unknown $g(x)$ has precisely the same form as the term $\sum_{i<t} s_w(X_i, X_t)\, Y_i / \sum_{i<t} s_w(X_i, X_t)$ appearing in equation (4). In our setup this term is part of the data-generating process. In addition, in nonparametric regression the bandwidth is selected so as to minimize some criterion, such as the mean integrated squared error, whereas we use maximum likelihood to estimate the weights w_j.

B. *Hypotheses Tests*

Rejecting the null hypothesis $H_0 : w_j = 0$ implies that the variable X_j contributes to the determination of Y, in the sense that the distance function, according to which Y_t is determined in equation (4), does not ignore the jth variable.

Under general conditions on the similarity function which are satisfied for exponential similarity, Lieberman (2005) proved that the MLE is weakly consistent is locally asymptotically mixed normal, and that

$$\sqrt{n}F^{-1/2}(\theta_0)[-P_n(\theta_0)](\hat{\theta}_n - \theta_0) \rightarrow_d N(0, I_{m+1}), \qquad (6)$$

where

$$F(\theta_0) = \lim_{n\to\infty} E_{\theta_0}\left(\frac{1}{n}\frac{\partial l(\theta_0)}{\partial\theta}\frac{\partial l(\theta_0)}{\partial\theta'}\right),$$

$$P_n(\theta) = \frac{1}{n}\frac{\partial^2 l(\theta)}{\partial\theta\,\partial\theta'},$$

and θ_0 is the true value of θ. For simple hypotheses tests for which the parameter is at the interior of the parameter space under the null, equation (6) can be used to apply any of the conventional likelihood-based tests (likelihood ratio, Lagrange multiplier, and Wald) in a straightforward manner. For a hypothesis of the form $H_0 : w_r = 0$ versus $w_r > 0$ ($r = 1, \ldots, m$), the parameter is on the boundary under H_0. For this case, Chant (1974, equation (8)) showed that the distribution of the normalized MLE is half-normal. Hence, for a one-sided t-test of the form above we reject H_0 as we do in the usual case when t is large (for example, when it exceeds 1.645, if a 5% significance level is desired).

V. Statistical Inference for the Discrete Case

The Model and the Likelihood Function

We now deal with the case in which each Y_t is categorical. In particular, consider examples 5 and 6 in section I A in which an expert is asked to estimate the probability of a certain event, and the observed values of Y can only be {0, 1}. A probability estimated by our formula with the empirical similarity function may be viewed as "objective" in that it does not rely on subjective similarity judgments, while still allowing different data points to have differing relevance to the estimation problem at hand.

When assessing probabilities, the assessed values can be anywhere in the interval [0, 1]. Indeed, formula (2) may generate any value in [0, 1]. But in this case one cannot assume that previously observed values of Y were generated by a normal distribution centered around a similarity-weighted average as in model (4).[9]

We therefore assume the following model:[10]

$$P(Y_t = 1 | \mathcal{F}_{t-1}) = F_t(z_t(w)), \qquad t = 1, \ldots, n, \qquad (7)$$

where F_t is a continuous conditional distribution function with density f_t, $\mathcal{F}_{t-1} = \sigma(Y_{t-1}, \ldots, Y_1; X_t, \ldots, X_1)$, and

$$z_t = \frac{\sum_{i<t} s_w(X_i, X_t)Y_i}{\sum_{i<t} s_w(X_i, X_t)}. \qquad (8)$$

In this setting the X's are taken to be fixed. Letting F_t be the standard normal cumulative distribution function (cdf) leads to a probit-type model whereas letting F_t be the logistic distribution leads to a logit-type model. Because $z_t \in [0, 1]$, it might be sensible to let F_t be a beta distribution or, quite simply, the uniform distribution. Note that in the classical model, it is postulated that $P(Y_t = 1|X) = F(X\beta)$. Where no Y_j's appear on the right-hand side, also observe that model (7) is nonlinear through both F_t and z_t.

In view of equations (7) and (8),

$$\frac{\partial P(Y_t = 1|\mathcal{F}_{t-1})}{\partial s_w(X_j, X_t)} = f_t(z_t)\frac{\sum_{i<t} s_w(X_i, X_t)(Y_j - Y_i)}{[\sum_{i<t} s_w(X_i, X_t)]^2},$$

which is non negative if $Y_j = 1$ and nonpositive if $Y_j = 0$. In other words, when the similarity between Y_t and Y_j increases, the conditional probability that $Y_t = 1$ will not fall when $Y_j = 1$ and will not rise when $Y_j = 0$.

Given our setup, the log likelihood is given by

$$l = \ln L = \sum_{t=1}^{n}\{Y_t \ln F_t(z_t) + (1 - Y_t)\ln[1 - F_t(z_t)]\},$$

[9] Other reasons for which the model (4) is inappropriate in this case are that the R^2 of regression would typically be low and that, because of the non-Gaussian nature of the observations, OLS would be inefficient.

[10] The categorical variables we discuss here may only assume the values 0 or 1. However, the analysis that follows can be extended to the case of a categorical variable assuming more than two categories.

and the MLE's are the solutions of

$$\frac{\partial l}{\partial w_j} = \sum_{t=1}^{n} \frac{Y_t - F_t}{F_t(1 - F_t)} f_t \upsilon_{t,j} = 0, \qquad j = 1, \ldots, m,$$

where

$$\begin{aligned} \upsilon_{t,j} &= \partial z_t / \partial w_j \\ &= \frac{[\sum_{i<t} \dot{s}_{w,j}(X_i, X_t)\, Y_i]\,[\sum_{i<t} s_w(X_i, X_t)]}{[\sum_{i<t} s_w(X_i, X_t)]^2} \\ &\quad - \frac{[\sum_{i<t} \dot{s}_{w,j}(X_i, X_t)]\,[\sum_{i<t} s_w(X_i, X_t)\, Y_i]}{[\sum_{i<t} s_w(X_i, X_t)]^2}, \end{aligned}$$

$$\dot{s}_{w,j}(X_i, X_t) = \frac{\partial s_{w,t,i}}{\partial w_j} = -\frac{s_{w,t,i}(X_{ji} - X_{jt})^2}{2d(X_i, X_j)},$$

and d is given in equation (3). As in the previous section, any likelihood-based procedure can be employed for hypothesis tests of the form H_0: $w = w_0$.

VI. Concluding Remarks

In the statistical analysis we assumed that each observation Y_t is distributed around a weighted average of past Y_i, or that $P(Y_t = 1)$ depends on such a weighted average. Such an ordering is necessary for a causal interpretation of our models. But if we consider a noncausal relationship, we may assume a model in which the distribution of each Y_i conditional on the other variables $\{Y_k\}_{k \neq i}$ is, say, normal around the weighted average of $\{Y_k\}_{k \neq i}$. Indeed, such a model may be more natural for applications in which the data are not naturally ordered. For this case, one should adapt the statistical model and the estimation of the similarity function accordingly.

The assumptions underlying our estimation process call for elaboration. The axiomatic model aims to describe how an *assessment* of Y_p, $\bar{Y}^s_p$, is generated based on *actually observed* values of the variable in question, namely, past values $(Y_i)_{i \leq n}$, such as selling prices of houses or of paintings. Applied to each past observation Y_i, it suggests that the *assessment* of Y_i, $\bar{Y}^s_i$, is generated by equation (2) for *actually observed* past values $(Y_k)_{k<i}$. That is, when we explain Y_i by past observations $(Y_k)_{k<i}$, we treat Y_i as if it were an assessment. When we explain Y_l for $l > i$, we treat Y_i as if it were an actual value. What justifies this confusion between the actual value of a variable and an assessment thereof?

For many applications of interest the answer lies in the notion of equilibrium. If all economic agents agree in their assessment of the price of a house or a painting, this joint assessment will indeed be its market price. Similarly, the price of a financial asset would equal its own assessment if all agents agreed on the latter. In these cases, one may assume that, as a feature of equilibrium, actually observed data coincide with their assessments.[11]

There may be applications in which one has direct access to, or indirect measurement of, both actual values (Y_i) and their assessments, say (Z_i). In these cases one may find the similarity function s that best fits the data according to

$$\bar{Z}^s_i = \frac{\sum_{k<i} s(X_k, X_i) Y_k}{\sum_{k<i} s(X_k, X_i)}, \tag{9}$$

namely, a function s that provides values $(\bar{Z}^s_i)_i$ that are close to $(Z_i)_i$, and then use this function to generate an estimate $\bar{Z}^s_p$ of Z_p using actual values Y_k by equation (9) applied to $i = p$.

Yet another class of applications involves only the assessments (Z_i). Assume, for instance, that one only observes asking prices (Z_i), and not actual selling prices (Y_i). [This is the case in Gayer, Gilboa, and Lieberman, 2004.] If everyone has access only to the asking prices (Z_i), one may apply our axiomatization to these variables, and conclude that the asking price of a new observation Z_p will be a similarity-weighted average of past asking prices $(Z_i)_{i \leq n}$. Moreover, it makes sense to assume that the same similarity function governed the generation of past values Z_i as a function of their past, $(Z_k)_{k<i}$. Hence one may estimate the similarity function in equation (2) with Z_i instead of Y_i, and use the estimated similarity for the prediction of Z_p.

Finally, in some situations one does not have access to the assessments (Z_i), and there is no theoretical reason to assume that $Z_i = Y_i$. In these cases our empirical approach could still be applied. That is, one may still ask which function $s : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_{++}$ can best fit the data under the assumption that they were generated by equation (2), and that function can then be used for prediction of Y_p by equation (1). In this type of application, (Y_i) can be viewed as proxies for (Z_i). Observe that it is only in the estimation of s that we replace (Z_i) by (Y_i). In the generation of the prediction $\bar{Z}^s_p$ using the estimated s, we use the actual values (Y_i), as indeed we should.

This paper is devoted to the theory of similarity-weighted averaging. This technique is used in Gayer et al. (2004) for the assessment of real estate prices, as in example 1. Their paper compared this method, representing case-based reasoning, with linear regression, representing rule-based reasoning.

[11] To a lesser degree, the rate of inflation and the probability of a stock market crash are also influenced by what economic agents assess them to be.

REFERENCES

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

Billot, A., I. Gilboa, D. Samet, and D. Schmeidler, "Probabilities as Similarity-Weighted Frequencies," *Econometrica* 73 (2005), 1125–1136.

Billot, A., I. Gilboa, and D. Schmeidler, "An Axiomatization of an Exponential Similarity Function," Cowles Foundation discussion paper no. 1485 (2004).

Chant, D, "On Asymptotic Tests of Composite Hypotheses in Non standard Conditions," *Biometrika* 61 (1974), 291–298.

Cover, T., and P. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory* 13 (1967), 21–27.

Devroye, L., L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition* (New York: Springer-Verlag, 1996).

Fix, E., and J. Hodges, "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties," Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX, technical report 4, (1951).

——— "Discriminatory Analysis: Small Sample Performance," USAF School of Aviation Medicine, Randolph Field, TX, technical report 21-49-004 (1952).

Gayer, G., I. Gilboa, and O. Lieberman, "Rule-Based and Case-Based Reasoning in Housing Prices," Cowles Foundation discussion paper no. 1493 (2004).

Gilboa, I., and D. Schmeidler, "Case-Based Decision Theory," *Quarterly Journal of Economics* 110 (1995), 605–639.

——— "Act Similarity in Case-Based Decision Theory," *Economic Theory* 9 (1997), 47–61.

——— *A Theory of Case-Based Decisions* (Cambridge: Cambridge University Press, 2001).

——— "Inductive Inference: An Axiomatic Approach," *Econometrica* 71 (2003), 1–26.

Hume, D., *Enquiry into the Human Understanding* (Oxford; Clarendon Press, 1748).

Lieberman, O., "Asymptotic Theory for Empirical Similarity Models," Technion mimeograph (2005).

Parzen, E. "On the Estimation of a Probability Density Function and the Mode," *Annals of Mathematical Statistics* 33 (1962) 1065–1076.

Riesbeck, C. K., and R. C. Schank, *Inside Case-Based Reasoning* (Hillsdale, NJ: Lawrence Erlbaum Associates, 1989).

Rosenblatt, M., "Remarks on Some Nonparametric Estimates of a Density Function," *Annals of Mathematical Statistics* 27 (1956), 832–837.

Schank, R. C., *Explanation Patterns: Understanding Mechanically and Creatively* (Hillsdale, NJ: Lawrence Erlbaum Associates, 1986).

Scott, D. W., *Multivariate Density Estimation: Theory, Practice, and Visualization* (New York: Wiley, 1992).

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

APPENDIX

Proofs

PROOF OF THEOREM 1: We begin by proving sufficiency of the axioms [that is, that (i) implies (ii)] and the uniqueness of the function s. Consider a pair $\xi, \eta \in \mathbb{R}$. Restricting $\{\succsim_I\}_{I\in\mathcal{I}}$ to $\{\xi, \eta\}$, one notices that they satisfy the conditions of the representation theorem in Gilboa and Schmeidler (2001, Theorem 3.1, p. 67, or 2003, Theorem 2, p. 16). Indeed, the first three axioms follow directly from A1–A3, whereas the diversity axiom for two alternatives follows from the averaging axiom, A4. To apply this theorem we have also to define the trivial relation for the memory $I = 0$: $\succsim_0 = \mathbb{R} \times \mathbb{R}$. Hence there exists a function $\upsilon^{\xi\eta}: \Gamma \times \mathbb{R} \to \mathbb{R}$, unique up to multiplication by a positive number, such that, for every $I \in \mathcal{I}$,

$$\xi \succsim_I \eta \quad \text{iff} \quad \sum_{(x,y)\in C} I((x,y))\upsilon^{\xi\eta}(x,y) \geq 0.$$

Next consider a triple $\{\xi, \eta, \zeta\} \subset \mathbb{R}$. Restricting $\{\succsim_I\}_{I\in\mathcal{I}}$ to $\{\xi,\eta,\zeta\}$ will no longer satisfy the diversity axiom in Gilboa and Schmeidler (2001). This axiom would state that for every permutation of the triple $\{\xi, \eta, \zeta\}$ there exists $I \in \mathcal{I}$ such that $\succ_I$ agrees with that permutation. This condition does not follow from our A4. Indeed, the diversity axiom is too strict for our purposes. If $\xi > \eta > \zeta$, then no $\succsim_I$ represented by (*) will satisfy $\xi \succ_I \zeta \succ_I \eta$.

However, the proof of Gilboa and Schmeidler's theorem does not require the full strength of their diversity axiom. All that is required for three alternatives $\{\xi, \eta, \zeta\}$ is that $\upsilon^{\xi\eta}$ not be a multiple of $\upsilon^{\eta\zeta}$. To this end, it suffices that there be three different permutations in $\{\succ_I\}_{I\in\mathcal{I}}$ (restricted to $\{\xi, \eta, \zeta\}$). This latter condition is guaranteed by our A4. Specifically, by the averaging axiom A4 for every distinct (ξ, η, ζ), there exists $I \in \mathcal{I}$ such that $\xi \succ_I \eta, \zeta$. Hence there are at least three permutations in $\{\succ_I\}_{I\in\mathcal{I}}$ (restricted to $\{\xi, \eta, \zeta\}$), and the representation theorem for triples holds. Observe also that this argument does not employ all the relations $\{\succsim_I\}_{I\in\mathcal{I}}$, and it can also be used for a restricted domain $\{\succsim_I\}_{I\in\mathcal{I}_x}$ for any $x \in \Gamma$.

It follows that, for every triple $\{\xi,\eta,\zeta\}$, one can find $\upsilon^{\xi},\upsilon^{\eta},\upsilon^{\zeta}: \Gamma \times \mathrm{R} \to \mathrm{R}$ such that, for every $a, b \in \{\xi, \eta, \zeta\}$ and for every $I \in \mathcal{I}$,

$$\begin{aligned} a \succsim_I b \quad \text{iff} \quad & \sum_{(x,y)\in C} I((x,y))\upsilon^{a}(x,y) \geq \sum_{(x,y)\in C} I((x,y))\upsilon^{b}(x,y) \\ \Leftrightarrow \quad & \sum_{(x,y)\in C} I((x,y))\,[\upsilon^{a}(x,y) - \upsilon^{b}(x,y)] \geq 0. \end{aligned} \tag{A-1}$$

In this case, the matrix $(\upsilon^{\xi}, \upsilon^{\eta}, \upsilon^{\zeta})$ is unique up to multiplication by a positive constant and addition of a constant to each row. Fix one such matrix $(\upsilon^{\xi}, \upsilon^{\eta}, \upsilon^{\zeta})$.

Fix $x \in \Gamma$ and consider $\mathcal{I}_x$. Restrict attention to $\{\succsim_I\}_{I\in\mathcal{I}_x}$. Because equation (A-1) applies to all $I \in \mathcal{I}$, it definitely holds for all $I \in \mathcal{I}_x \subset \mathcal{I}$. However, we claim that, even on this restricted domain, the matrix $(\upsilon^{\xi}, \upsilon^{\eta}, \upsilon^{\zeta})$ is unique as above. To see this, recall that our derivation of equation (A-1) coupled with the uniqueness result, holds true for $\{\succsim_I\}_{I\in\mathcal{I}_x}$ for any $x \in \Gamma$.

Observe that the relations $\{\succsim_I\}_{I\in\mathcal{I}_x}$ are completely specified by A4. Specifically, for every $a, b \in \mathbb{R}$, for every $I \in \mathcal{I}_x$,

$$a \succsim_I b \quad \text{iff} \quad |a - y_I| \leq |b - y_I|, \tag{A-2}$$

where

$$y_I = \frac{\sum_{(x,y)\in C} I((x,y))\,y}{\sum_{(x,y)\in C} I((x,y))}.$$

That is, $a \succsim_I b$ iff a is closer to the average y_I than is b. Consider

$$f_I(\alpha) = \sum_{(x,y)\in C} I((x,y))(\alpha - y)^2.$$

The function $f_I(\alpha)$ is quadratic (in α), and it has a minimum at $\alpha = y_I$. It follows that for every a, b, for every $I \in \mathcal{I}_x$,

$$f_I(a) \leq f_I(b) \quad \text{iff} \quad |a - y_I| \leq |b - y_I|.$$

Combining this fact with the definition of f_I and with equation (A-2), we conclude that, for every $a, b \in \{\xi, \eta, \zeta\}$ and for every $I \in \mathcal{I}_x$

$$\begin{aligned} a \succsim_I b \quad \text{iff} \quad & \sum_{(x,y)\in C} I((x,y))(a - y)^2 \leq \sum_{(x,y)\in C} I((x,y))(b - y)^2 \\ \Leftrightarrow \quad & \sum_{(x,y)\in C} I((x,y))[(a - y)^2 - (b - y)^2] \leq 0. \end{aligned} \tag{A-3}$$

The uniqueness of the representation in equations (A-1) and (A-3) imply that there exists a constant $s(x) > 0$ such that

$$\upsilon^{a}(x, y) - \upsilon^{b}(x, y) = -s(x)[(a - y)^2 - (b - y)^2] \tag{A-4}$$

for every $a,b \in \{\xi, \eta, \zeta\}$ and for every $y \in \mathrm{R}$. Obviously, once $\upsilon^{a}(x,y)$, $\upsilon^{b}(x, y)$ are fixed, $s(x)$ is uniquely determined by equation (A-4).

We now turn to discuss various x's, while still focusing on the triple $\{\xi,\eta, \zeta\}$. Consider I in $\mathcal{I}$ (but not necessarily in $\mathcal{I}_x$ for any x). Combine equations (A-1) and (A-4) to conclude that, for $a,b \in \{\xi, \eta, \zeta\}$ and for all $I \in \mathcal{I}$,

$$\begin{aligned} a \succsim_I b \quad \text{iff} \quad & \sum_{(x,y)\in C} I((x,y))s(x)(a - y)^2 \\ & \leq \sum_{(x,y)\in C} I((x,y))s(x)(b - y)^2 \\ \Leftrightarrow \quad & \sum_{(x,y)\in C} I((x,y))s(x)[(a - y)^2 - (b - y)^2] \leq 0. \end{aligned} \tag{A-5}$$

Define

$$g_I(\alpha) = \sum_{(x,y)\in C} I((x,y))s(x)(a-y)^2.$$

Like the function f_I above, the function $g_I(\alpha)$ is quadratic (in α), and it has a minimum at

$$\alpha = y_{s,I} = \frac{\sum_{(x,y)\in C} s(x)I((x,y))\,y}{\sum_{(x,y)\in C} s(x)I((x,y))}.$$

It follows that for every a, b, for every $I \in \mathcal{I}$,

$$g_I(a) \le g_I(b) \quad \text{iff} \quad |a - y_{s,I}| \le |b - y_{s,I}|. \tag{A-6}$$

Combining equation (A-5) with (A-6), we obtain

$$a \gtrsim_I b \quad \text{iff} \quad |a - y_{s,I}| \le |b - y_{s,I}|,$$

that is, $\{s(x)\}_x$ satisfies (∗) for the triple $\{\xi,\eta,\zeta\}$.

Observe that $\{s(x)\}_x$ are unique up to multiplication by a positive number. In fact, we argue that if s and s' both satisfy (∗) for particular a, $b \in$ R, $a \neq b$, then there exists $\lambda > 0$ such that $s'(x) = \lambda s(x)$ for all $x \in \Gamma$. Indeed, assume that s and s' both satisfy (∗) for particular a, b. This would imply that they both satisfy equation (A-5) for these a, b, and then the uniqueness of $\upsilon^a(x, y) - \upsilon^b(x, y)$ in equation (A-1), combined with equation (A-4), implies that there exists $\lambda > 0$ such that $s'(x) = \lambda s(x)$ for all $x \in \Gamma$.

It remains to show that the function $s(x)$ does not depend on the choice of the triple $\{\xi,\eta,\zeta\} \subset \mathbb{R}$. Consider the triple $\{\xi, \eta, \tau\}$ where $\tau \neq \zeta$. Because equation (A-6) applied to ξ and η holds both for the function s of the triple $\{\xi,\eta, \zeta\}$ and that of the triple $\{\xi, \eta, \tau\}$ these two functions have to be positive multiples of each other. Using this argument inductively implies that all functions s derived from different triples differ only by a factor. Because s can always be multiplied by a positive constant and still satisfy (∗), one may choose an s of one triple $\{\xi,\eta, \zeta\}$ arbitrarily and use it for all other triples as well.

We need to prove the necessity of the axioms, that is, that (ii) implies (i). The necessity of A1, A2, and A3 is proved as in Gilboa and Schmeidler (2001, 2003), whereas the necessity of A4 follows directly from (∗).

Proof of Corollary 2: This result is a rewriting of theorem 1 for the special case in which C is the product of two sets.

Proof of Corollary 3: Use corollary 2, and define $s(x) = \frac{1}{p}\sum_{i\in P} s_i(x)$. For $I \in \mathcal{I}_{\text{sh}}$,

$$\frac{\sum_{(i,x,y)\in C} s_i(x)I((i,x,y))\,y}{\sum_{(i,x,y)\in C} s_i(x)I((i,x,y))} = \frac{\sum_{(i,x,y)\in C} s(x)I((i,x,y))\,y}{\sum_{(i,x,y)\in C} s(x)I((i,x,y))} = y_{s,I},$$

which concludes the proof.