Journal of Econometrics 162 (2011) 124–131

Contents lists available at ScienceDirect

Journal of Econometrics

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

A similarity-based approach to prediction☆

Itzhak Gilboa[a,b,c,*], Offer Lieberman[d], David Schmeidler[a,e]

[a] Tel-Aviv University, Israel
[b] HEC, Paris, France
[c] Cowles Foundation, Yale University, USA
[d] University of Haifa, Israel
[e] The Ohio State University, USA

ARTICLE INFO

Article history:
Available online 29 October 2009

Keywords:
Density estimation
Empirical similarity
Kernel
Spatial models

ABSTRACT

Assume we are asked to predict a real-valued variable y_t based on certain characteristics $x_t = (x_t^1, \ldots, x_t^d)$, and on a database consisting of $(x_i^1, \ldots, x_i^d, y_i)$ for $i = 1, \ldots, n$. Analogical reasoning suggests to combine past observations of x and y with the current values of x to generate an assessment of y by *similarity-weighted averaging*. Specifically, the predicted value of y, y_t^s, is 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 $x_t^1, \ldots, x_t^d$, associated with y_t, and the previously observed vector, $x_i^1, \ldots, x_i^d$. The "empirical similarity" approach suggests estimation of the similarity function from past data. We discuss this approach as a statistical method of prediction, study its relationship to the statistical literature, and extend it to the estimation of probabilities and of density functions.

1. Introduction

Reasoning by analogies is a basic method of predicting future events based on past experience. Hume (1748), who famously questioned the logical validity of inductive reasoning, also argued that analogical reasoning is the fundamental tool by which we learn from the past about the future. Analogical reasoning has been widely studied in psychology and artificial intelligence (see Schank, 1986; Riesbeck and Schank, 1989), and it is very common in everyday discussions of political and economic issues. Furthermore, it is a standard approach to teaching in various professional domains such as medicine, law, and business. However, analogical reasoning has not been explicitly applied to statistics. The goal of this paper is to present an analogy-based statistical method, and to explore its relationships to existing statistical techniques.

Suppose that we are trying to assess the value of a variable y_t based on the values of relevant variables, $x_t = (x_t^1, \ldots, x_t^d)$, and on a database consisting of the variables $(x_i^1, \ldots, x_i^d, y_i)$ for $i = 1, \ldots, n$. For example, y_t may be the price of an antique piece of furniture, where x_t denotes certain characteristics thereof, such as its style, period, size, and so forth. Alternatively, y_t may be an indicator variable, denoting whether a PhD candidate completes her studies successfully, where x_t specifies what is known about the candidate at the time of admission, including such variables as GRE and GPE scores, the ranking of the college from which the candidate graduated, etc.

How should we combine past observations of x and y with the current values of x to generate an assessment of y? If we were to follow Hume's idea, we would need a notion of similarity, indicating which past conditions $x_i = (x_i^1, \ldots, x_i^d)$ were more similar and which x_i's were less similar to x_t. We would like to give the observations that were obtained under more similar conditions a higher weight in the prediction of y_t than those who were obtained under less similar conditions. In the examples above, it makes sense to assess the price of an antique by the price of other, similar antiques that have recently been sold. Moreover, the more similar is a previous observation to the current one – in terms of style, period, size, and even time of sale – the greater is the weight we would like to put on this observation in the current assessment. Similarly, in assessing the probability of success of a PhD candidate, it seems desirable to put more weight on the observed outcomes involving more similar candidates as compared to less similar ones.

In attempting to let previous cases matter for a current prediction problem, but to do so in varying degrees, a similarity-weighted average is arguably the most natural formula. Formally, one may assume that there is a similarity function $s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{++} =$

☆ We are grateful to two anonymous referees for their comments. We also gratefully acknowledge financial support from Israel Science Foundation Grant No. 355/06. Gilboa and Schmeidler also acknowledge support from the Pinhas Sapir Center for Development and from the Polarization and Conflict Project CIT-2-CT-2004-506084 funded by the European Commission-DG Research Sixth Framework Programme.

* Corresponding address: Tel-Aviv University, Tel-Aviv 69978, Israel.
E-mail addresses: tzachigilboa@gmail.com (I. Gilboa), offerl@econ.haifa.ac.il (O. Lieberman), schmeid@tau.ac.il (D. Schmeidler).

0304-4076/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2009.10.015

$(0, \infty)$ such that, given a database $(x_i, y_i)_{i\le n}$ and a new data point $x_t = (x_t^1, \ldots, x_t^d) \in \mathbb{R}^d$, the similarity-based predictor of y_t is

$$y_t^s = \frac{\sum_{i\le n} s(x_i, x_t) y_i}{\sum_{i\le n} s(x_i, x_t)}. \tag{1}$$

Observe that, in the case when all similarity values are constant, this formula boils down to a simple average of past observations. The sample average is arguably the most basic and most widely used statistic. As such, the formula (1) appears to be a minor variation on the averaging principle. Rather than a simple average, we suggest using a weighted one, where the weights reflect the relevant similarity. If we consider a limiting case where the function s is the indicator $s^*(x_i, x_t) = 1$ if $x_i = x_t$ and $s^*(x_i, x_t) = 0$ otherwise, (1) becomes the conditional sample average of y, given that $x = x_t$. Thus, (1) may be viewed as a continuous family of formulae spanning the range between the conditional and the unconditional average of past observations.

However, formula (1) is not the only way to simultaneously generalize averaging and conditional averaging. Is it more or less reasonable than others? What properties does it have? Such questions call for an axiomatic treatment.

Gilboa and Schmeidler (1995, 2001) suggested an axiomatic theory of case-based decision making. Gilboa and Schmeidler (2003) specialized the general theory to prediction problems. Their approach studies the way that possible predictions are ranked, as a function of the database of given observations. A key axiom in this paper is the so-called combination axiom, stating that a ranking that follows from two disjoint databases should also follow from their union. The main result uses the combination axiom, coupled with a few other axioms, to characterize a general prediction rule. It turns out that several statistical techniques are special cases of this general rule. In particular, kernel estimation of a density function, kernel classification, and maximum likelihood estimation are such special cases.

The axiomatic approach to statistical problems allows one to study the properties that characterize various techniques, to ask how reasonable these techniques are, and to find similarities between them. For example, Gilboa and Schmeidler (2003) discuss the combination axiom and attempt to come up with general guidelines for the classification of applications in which it may be reasonable. Such a discussion may enrich our understanding of the statistical techniques that satisfy this axiom. Moreover, the axiomatic treatment exposes similarities that may not be otherwise obvious, such as the similarity between kernel classification and maximum likelihood estimation. At the same time, the axiomatic analysis also makes it easier to come up with "counter-examples", that is, with situations in which axioms are implausible, thereby delineating the scope of applicability of various techniques. In particular, the combination axiom appears less compelling for time series than it is for cross-sectional datasets. Correspondingly, applying formula (1) where t denotes time may be inappropriate.[1] We maintain that the axiomatic approach may benefit statistical theory in general, because axioms may be viewed as criteria for the evaluation of statistical techniques in finite samples.

Applying the axiomatic approach to the problem at hand, Gilboa, Lieberman, and Schmeidler (GLS, 2006) axiomatized formula (1) for the case that y is a real-valued variable, while Billot, Gilboa, Samet, and Schmeidler (BGSS, 2005) axiomatized it for the case that y is a multi-dimensional probability vector. These papers do not assume that the similarity function is given. Rather, they consider a certain observable measure – such as a likelihood ordering or a probability assessment – and ask how this observable measure varies with the database that is the input to the problem. The axiomatizations impose certain constraints on the way the observable measure varies with the input database, and prove that the constraints are satisfied if and only if there exists a similarity function such that (1) holds.

The formula (1) may be used with any function $s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{++}$. Which function should we choose? GLS (2006) suggest obtaining the similarity function from the data, selecting the function s that best fits the data. The notion of "best fit" can be defined within a statistical model or otherwise. A non-statistical approach, often used in machine learning, does not specify a data generating process (DGP). Rather, it selects a best-fit criterion such as minimal sum of squared errors. Alternatively, the formula (1) can be embedded within a statistical model, parametric or non-parametric. In either case, the optimal s is computed from the data. (See details in Section 2 below.)

The right-hand side of formula (1) is mathematically equivalent to a kernel estimator of a non-parametric function, where the similarity function plays the role of the kernel. Thus, the axiomatic derivations of this formula in GLS (2006) and BGSS (2005) may be viewed as axiomatizing kernel-based non-parametric methods. If one takes GLS (2006) and BGSS (2005) as a descriptive model of human reasoning, one might argue that the Nadaraya–Watson estimator of an unknown function coincides with the way the human mind has evolved to predict variables. Indeed, since the human mind is supposed to be a general inference tool, capable of making predictions in unknown environments, it stands to reason that it solves a non-parametric statistical prediction problem.

The main contributions of the present paper are to relate the empirical similarity approach to the statistical literature, and to extend it to the problem of density estimation, where the density of a variable y_t is assumed to depend on observable variables $x_t = (x_t^1, \ldots, x_t^d)$.

Section 2 describes the empirical similarity statistical models. We devote Section 3 to a more detailed discussion of the relationship between kernel-based estimation and empirical similarity. We then briefly discuss the relationship of our method to spatial models in Section 4. Section 5 discusses the case of a binary random variable. In Section 6 we apply our method to the non-parametric estimation of a density function, and provide an axiomatization of a "double-kernel" estimation method. Finally, Section 7 concludes with a discussion of additional directions for future research.

2. Empirical similarity models

Which function $s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{++}$ best explains the database $(x_i, y_i)_{i\le n}$? This question, which may or may not be couched in a statistical model, would take a different form depending on whether the data are naturally ordered. If they are, such that for every $i > j$, (x_i, y_i) was realized after (x_j, y_j), it is natural to consider the similarity-based predictor of y_i, for a given s, to be

$$y_i^s = \frac{\sum_{j<i} s(x_j, x_i) y_j}{\sum_{j<i} s(x_j, x_i)}. \tag{2}$$

If, however, the order of the datapoints in $(x_i, y_i)_{i\le n}$ is arbitrary, it is more natural to define

$$y_i^s = \frac{\sum_{j\ne i} s(x_j, x_i) y_j}{\sum_{j\ne i} s(x_j, x_i)}. \tag{3}$$

In either case, the choice of the function s may be guided partly by theoretical considerations. Billot et al. (2008) provide

[1] In GLS (2006) we suggest that time series may be analyzed by defining similarities over patterns, or subsequences of observations.

conditions on similarity-weighted averages that are equivalent to the similarity function taking the form

$$s(x, x') = \exp(-\|x - x'\|)$$

where $\|\cdot\|$ is a norm on $\mathbb{R}^d$. For concreteness, we focus on the family of norms defined by weighted Euclidean distances.

$$s_w(x, x') = \exp\left(-d_w(x, x')\right)$$

where $w \in \mathbb{R}^d_+$ is a weight vector such that the distance between two vectors $x, x' \in \mathbb{R}^d$ is given by

$$d_w(x, x') = \sum_{j=1}^{d} w_j \left(x_j - x'_j\right)^2. \tag{4}$$

Thus, the similarity function is known up to a d-dimensional vector of parameters, one for each predictor.

In order to conduct statistical inference and to obtain qualitative results by hypotheses tests, one may embed Eqs. (2) and (3) within a statistical model, namely

$$y_t = \frac{\sum_{i<t} s_w(x_i, x_t) y_i}{\sum_{i<t} s_w(x_i, x_t)} + \varepsilon_t, \tag{5}$$

and

$$y_t = \frac{\sum_{i\neq t} s_w(x_i, x_t) y_i}{\sum_{i\neq t} s_w(x_i, x_t)} + \varepsilon_t, \tag{6}$$

respectively, where $\{\varepsilon_t\}$ are iid $\left(0, \sigma^2\right)$.

Model (5) can be interpreted as an explicit causal model. Consider, for example, a process of price formation by case-based economic agents. These agents determine the prices of unique goods such as apartments or art pieces according to the similarity of these goods to other goods, whose prices have already been determined in the past.[2] Thus, (5) can be thought of as a model of the mental process that economic agents engage in when determining prices. The estimation of s_w in such a model is thus an estimation of a similarity function that presumably causally determines the observed y's. The asymptotic theory for this model was developed by Lieberman (in press).

Model (6) cannot be directly interpreted in the same way. Because the distribution of each y_t depends on all the other y_i's, (6) cannot be a temporal account of the evolution of the process. However, such interdependencies may be quite natural in geographical, sociological, or political data, as is common in spatial statistics (see Section 4 below).

Both models (5) and (6) assume that the similarity function is fixed and does not change with the realizations of y_t, nor with t itself. They rely on the axiomatizations in GLS (2006) and in BGSS (2005). Each of these axiomatizations, like Gilboa and Schmeidler (2001, 2003), uses a so-called "combination" (or "concatenation") axiom.[3] Whereas axioms of this type may appear reasonable at first, they are rather restrictive. Gilboa and Schmeidler (2003) contains an extensive discussion of such an axiom and its limitations, and the latter apply to all versions of the axiom, including those that appear in GLS (2006) and in BGSS (2005). For our purposes, it is important to note that the combination axiom does not allow one to learn the similarity function from the data. Correspondingly, formula (1) does not allow the similarity function to change with the accumulation of data. But the basic idea of "empirical similarity" is precisely this, namely, that the similarity function be learnt from the same data that are used, in conjunction with this similarity function, for generating predictions. Hence, the axiomatic derivations mentioned above are limited. Similarly, formula (1) calls for a generalization that would allow it to refine the similarity assessment, and the statistical models (5) and (6) should be accordingly generalized.

3. Empirical similarity and kernel-based methods

For clarity of exposition, we start with the unidimensional case, that is, when $d = 1$ and there is only one explanatory variable X. A nonparametric regression model assumes a DGP of the following type:

$$y_i = m(x_i) + \varepsilon_i, \quad (i = 1, \ldots, n), \tag{7}$$
$$\varepsilon_i \sim iid\ \left(0, \sigma^2\right),$$

where x_i is a scalar and $m : \mathbb{R} \to \mathbb{R}$ is the unknown function relating x to y. A widely used nonparametric estimator of $m(\cdot)$ is the Nadaraya–Watson estimator, defined as

$$\hat{m}(x_t) = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x_t}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_i - x_t}{h}\right)}, \tag{8}$$

where $K(x)$ is a kernel function, that is, a non-negative function satisfying $\int K(z)\,dz = 1$, as well as other regularity conditions, and h is a bandwidth parameter. For instance, if we choose the Gaussian kernel, then

$$\frac{1}{h} K\left(\frac{x_i - x_t}{h}\right) = \left(2\pi h^2\right)^{-1/2} \exp\left(-\frac{(x_i - x_t)^2}{2h^2}\right). \tag{9}$$

The choice of h is central in the nonparametric literature, because there is a trade-off between variance and bias. One of the most common criteria for the selection of an optimal bandwidth is to minimize the mean integrated squared error (MISE). That is, the optimal h satisfies

$$h^* = \arg\min_h E_{f_0} \int \left(\hat{m}(x) - m(x)\right)^2 dx, \tag{10}$$

where the expectation is taken under the true density f_0 of y. If x is countable and $m(x)$ is replaced by y, then we end up with a minimum expected sum of squared errors criterion.

We now turn to discuss the connection between kernel-based estimation and empirical similarity. As described above, the empirical similarity method suggests predicting y_t by

$$y_t = \frac{\sum_{i=1}^{n} s_w(x_i, x_t)\, y_i}{\sum_{i=1}^{n} s_w(x_i, x_t)},$$

where

$$\begin{aligned} s_w(x_i, x_t) &= \exp(-d_w) \\ &= (\pi/w)^{1/2} \left[\frac{1}{\left(1/\sqrt{2w}\right)} K\left(\frac{x_i - x_t}{1/\sqrt{2w}}\right)\right], \end{aligned}$$

d_w was defined in (4), and K is given in (9). Then,

$$\frac{\sum_{i=1}^{n} s_w(x_i, x_t)\, y_i}{\sum_{i=1}^{n} s_w(x_i, x_t)} = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x_t}{1/\sqrt{2w}}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_i - x_t}{1/\sqrt{2w}}\right)}.$$

[2] See Gayer et al. (2007).

[3] A variant of this axiom is also used in the axiomatization in Section 6.

It follows that, in this setting,

$$h = 1/\sqrt{2w}.$$

Thus, we have a direct mapping from the similarity parameter to the bandwidth parameter. Among other things, we can set w^* to satisfy the MISE criterion.

Despite the similarity between kernel-based estimation and empirical similarity, there is a fundamental difference between them. The former is a statistical technique that is used, among other things, for the estimation of model (7). By contrast, in models (5) and (6) we use the formula (1) as part of the DGP itself.

This difference is accentuated when we focus on the ordered case. We can rewrite model (5) as

$$y_j = \hat{m}^w_{(j-1)}\left(x_j\right) + \varepsilon_j, \quad (j = 2, \ldots, n), \tag{11}$$

where $\hat{m}^w_{(j-1)}\left(x_j\right)$ is defined as in (8), restricted to the observations that precede j, namely

$$\hat{m}^w_{(j-1)}\left(x_j\right) = \frac{\sum_{i=1}^{j-1} s_w\left(x_i, x_j\right) y_i}{\sum_{i=1}^{j-1} s_w\left(x_i, x_j\right)}. \tag{12}$$

Model (7) assumes that the distribution of y_t is a function of x_t alone. If the function m were known, the best predictor of y_t given x_t would have been $m(x_t)$, independent of previous realizations of x and of y. In other words, model (7) specifies a *rule*, m, relating x_t to y_t. This is not the case for model (5). In this model, the DGP is *case* based, where the distribution of y_t depends on all past and present realizations of x, as well as all past realizations of y.

Observe that this difference also has an implication regarding the type of questions that are raised about the parameters w or h. In (7), the parameter h is chosen optimally, so as to minimize an expected loss function. It has a purely statistical purpose and meaning. But in (5) and (6), w has a model meaning. Similar to a regression parameter, w may have an economic, psychological, or other substantial meaning having to do with the interpretation of the model. Indeed, in GLS we develop tests for hypotheses of the form[4]

$$H_0 : w = 0.$$

That is, in this model "What is the true value of w?" is a meaningful question, whereas in (7) one may only ask "What is a useful value of h?".

Despite these differences, the mathematical connections established above suggest that one may also use the empirical similarity approach to predict the value of y even though, in reality, the true DGP is (7). One would then expect the empirical similarity function to become "tighter" with an increase in the database size. To consider an extreme example, assume that a database is replicated in precisely the same way a large number of times. For every past observation (x_i, y_i) there will be many identical observations, and the similarity function that best explains existing data will be one with infinite w, that is, a similarity function that ignores all but the identical x values.[5]

[4] Under the hypothesis that $w = 0$, $S_w\left(x_i, x_j\right) = 1$ for all i and j. This suggests that y is not influenced by x — past values of y are relevant to its current evaluation irrespective of the x values that were associated with them. Mathematically, setting $w = 0$ yields the same prediction as using a kernel approach with $h = \infty$, where for every x, y is evaluated by a simple average of all past y's.

[5] In fact, two replications would suffice for the above argument. But a large number of replications would have a similar impact even if the database is not replicated in precisely the same way.

The discussion above generalizes to higher dimensions ($d > 1$) without any fundamental modifications. Kernel estimation is used for estimation of a non-parametric model (7) where x is multi-dimensional, and the models (5) and (6) have also been formulated for a multi-dimensional x. Indeed, similar relationships exist between the kernel bandwidth parameters and the weights that determine the similarity function. Specifically, we may specify

$$\begin{aligned} s_w\left(x_i, x_t\right) &= \exp\left(-d_w\right) \\ &= (2\pi)^{d/2}\left(\det\left(W\right)\right)^{1/2}\left[\left(\det\left(W\right)\right)^{-1/2} K\left(x_i - x_t; W\right)\right], \end{aligned} \tag{13}$$

where W^{-1} is a diagonal matrix with elements $2w_j, j = 1, \ldots, d$, and the term in the square brackets of (13) integrates to one. In this setting

$$\frac{\sum_{i=1}^{n} s_w\left(x_i, x_t\right) y_i}{\sum_{i=1}^{n} s_w\left(x_i, x_t\right)} = \frac{\sum_{i=1}^{n} K\left(x_i - x_t; W\right) y_i}{\sum_{i=1}^{n} K\left(x_i - x_t; W\right)},$$

where the jth bandwidth h_j is equal to $1/\sqrt{2w_j}$.

The bulk of the literature on multivariate kernels focuses only on one bandwidth parameter, but there is no conceptual difficulty in optimizing a multi-dimensional bandwidth. This, indeed, has been discussed by Yang and Tchernig (1999). As in the univariate case, we find the same conceptual differences between the empirical similarity model and kernel estimation. In particular, the empirical similarity model allows one to test hypotheses of the form

$$H_0 : w_j = 0$$

suggesting that variable x_j is immaterial in similarity judgments. Rejecting such a hypothesis constitutes a statistical proof that the variable x_j matters for the assessment of y. By contrast, a kernel function that is not part of the DGP does not render itself to the testing of similar qualitative hypotheses.

4. Empirical similarity and spatial models

The general spatial model can be written in at least two ways, in each case leading to a different likelihood. Besag (1974, p. 201; see also Cressie, 1993) describes the two possibilities. First, the conditional density of y_i given $y_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$ is specified as

$$p_i\left(y_i | y_{-i}\right) = \left(2\pi\sigma^2\right)^{-1/2} \times \exp\left[-\frac{1}{2\sigma^2}\left\{y_i - \mu_i - \sum_{j\neq i}\beta_{i,j}\left(y_j - \mu_j\right)\right\}^2\right].$$

This results in the following joint density of $y = (y_1, \ldots, y_n)$:

$$p\left(y\right) = \left(2\pi\sigma^2\right)^{-n/2}\left|B\right|^{1/2}\exp\left[-\frac{1}{2\sigma^2}\left(y-\mu\right)' B\left(y-\mu\right)\right],$$

where $[B]_{i,i} = 1$, $[B]_{i,j} = [B]_{j,i} = -\beta_{ij}$ and B is positive definite. Alternatively, one can assume that

$$E\left(y_i | y_{-i}\right) = \mu_i + \sum_{j\neq i}\beta_{i,j}\left(y_j - \mu_j\right).$$

For example, this holds for the model

$$y_i = \mu_i + \sum_{j\neq i}\beta_{i,j}\left(y_j - \mu_j\right) + \varepsilon_i,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are iid normal variables with zero mean and variance σ^2. In this case the joint density is

$$p\left(y\right) = \left(2\pi\sigma^2\right)^{-n/2}\left|B\right|\exp\left[-\frac{1}{2\sigma^2}\left(y-\mu\right)' B'B\left(y-\mu\right)\right]. \tag{14}$$

It is required that B is positive definite. Note that if we define

$$[B]_{i,j} = -\frac{s_w(x_i, x_j)}{\sum_{j \neq i} s_w(x_i, x_j)},$$

then (14) is the joint density of y in the similarity model (6). This model is also entitled *conditional autoregression* (or CAR).

These spatial models resemble models (5) and (6). The latter may appear more restrictive than the spatial model, because the similarity function s_w specifies a particular functional form for the coefficients $\beta_{i,j}$ (and, in (5), there are additional constraints that $\beta_{i,j} = 0$ for $i < j$). However, in most spatial applications (e.g., Anselin, 1988) the $\beta_{i,j}$'s are taken to be fixed and given whereas in models (5) and (6) the coefficients are not assumed known. Rather, they are functions of the x's and the w's and therefore, they are ultimately estimated from the data.

5. Probability estimation

GLS (2006) also propose using the empirical similarity approach for the estimation of probabilities. Such probabilities may be used in a decision problem, employing expected utility maximization or some other decision procedure that is probability based, such as median-utility maximization. Our focus at this point is on probabilities *per se*.

In this context, consider $y_t \in \{0, 1\}$, as in the example of success in a PhD program mentioned above. GLS develop the likelihood function for the ordered model, in which the probability that $y_t = 1$ depends only on past observations, y_i for $i < t$, and this probability is taken to be the similarity-weighted average of these past observations, namely, the similarity-weighted frequency of 1's in the past[6]:

$$p^s_w(y_t = 1 | x_1, \ldots, x_t, y_1, \ldots, y_{t-1}) = \frac{\sum_{i<t} s_w(x_i, x_t)\, y_i}{\sum_{i<t} s_w(x_i, x_t)}. \tag{15}$$

However, there are many applications in which the given data are not ordered in any natural way. In this case, one may assume that the probability that each data point y_t, $t = 1, \ldots, n$, equals 1 is given by

$$p^s_w(y_t = 1 | x_1, \ldots, x_n, y_1, \ldots, y_{t-1}, y_{t+1}, \ldots, y_n) = \frac{\sum_{i \neq t} s_w(x_i, x_t)\, y_i}{\sum_{i \neq t} s_w(x_i, x_t)}. \tag{16}$$

If $p(y_i) = p$ for all i, then $p^s_w(y_t = 1|\cdot)$ is evidently unbiased for p. To estimate w, we can use the idea of likelihood cross-validation, as follows. First, we define

$$p^s_{w,-i}(y_i = 1 | x_1, \ldots, x_n, y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n) = \frac{\sum_{j \neq i} s_w(x_j, x_i)\, y_j}{\sum_{j \neq i} s_w(x_j, x_i)},$$

for $i, j = 1, \ldots, n$, which is the leave-y_i-out cross-validation first step. At the second stage of the procedure we obtain

$$\hat{w}_{CV} = \arg\max_w \sum_{i=1}^{n} \log\left(p^s_{w,-i}(y_i = 1 | x_1, \ldots, x_n, y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)\right).$$

Finally, for a new data point $t = n + 1$, we estimate (15) by

$$\hat{p}_{\hat{w}_{CV}}(y_t = 1 | x_1, \ldots, x_n, x_t, y_1, \ldots, y_n) = \frac{\sum_{i=1}^{n} s_{\hat{w}_{CV}}(x_i, x_t)\, y_i}{\sum_{i=1}^{n} s_{\hat{w}_{CV}}(x_i, x_t)}.$$

Note the difference between this procedure and the one discussed in Silverman (1986, pp. 126–127). In our notation, Silverman's equation (6.7) reduces to

$$\hat{p}(y_t = 1) = \frac{\lambda}{n} \sum_{i=1}^{n} \left(\frac{1-\lambda}{\lambda}\right)^{(y_i - y_t)^2}, \tag{17}$$

where λ is a parameter, assumed to lie in $[1/2, 1]$, to be estimated by likelihood cross-validation. That is,

$$\hat{\lambda}_{CV} = \arg\max_{\lambda} \sum_{i=1}^{n} \log\left(\hat{p}_{-i}(y_i = 1)\right)$$

with

$$\hat{p}_{-i}(y_i = 1) = \frac{\lambda}{n} \sum_{j \neq i} \left(\frac{1-\lambda}{\lambda}\right)^{(y_j - y_i)^2}.$$

Unlike the case of nonparametric estimation of $m(x)$ with unordered data, it is not apparent how we can map λ into w. Also, with the "right" choice of s_w it might be possible to find a similarity-based predicted probability which outperforms (17) in terms of the sum of squared errors.

6. Double kernel density estimation

Suppose that one wishes to estimate the density function of a real-valued variable y, where this density is assumed to depend on the values of other real-valued variables $x = (x^1, \ldots, x^d)$. Assume that the jth past observation is a vector $(x^1_j, \ldots, x^d_j, y_j) \in \mathbb{R}^{d+1}$, $j = 1, \ldots, t-1$. A new datapoint $x_t \in \mathbb{R}^d$ is given. How should we estimate the density of y given x_t?

Kernel estimation of a density function is a well-known and widely used technique for the case in which there are no explanatory variables $x^1, \ldots, x^d$. (See Akaike, 1954; Rosenblatt, 1956; Parzen, 1962; Silverman, 1986; Scott, 1992.) It is therefore a natural candidate for a starting point. One can therefore ask a more concrete question: How can we generalize kernel estimation to the current problem, in which the density of y is assumed to depend on the realization of the variables $x^1, \ldots, x^d$?

Gilboa and Schmeidler (2003) used a "combination" axiom to derive kernel estimation of a density function for the standard case, in which there are no explanatory variables. As mentioned above, variants of this combination axiom are at the heart of the derivation of the similarity-weighted averages in BGSS (2005) and GLS (2006). It therefore appears coherent to estimate the density of y by a kernel method, but to allow the kernel to depend on the explanatory variables $x^1, \ldots, x^d$ in a way that resembles the similarity-weighted average used above.

Specifically, assume that there exists a function $s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{++}$, where $s(x_t, x_j)$ measures the degree to which data point $x_t \in \mathbb{R}^d$ is similar to data point $x_j \in \mathbb{R}^d$, and a kernel function

[6] GLS also allow the probability to depend on this similarity-weighted frequency in a monotone way. The more specific assumption, namely, that the similarity-weighted frequency *is* the probability, suggests an interpretation of "probability" that generalizes the frequentist definition, while retaining its intuitive appeal. However, this model cannot describe how the process starts and generates both 0's and 1's.

$K : \mathbb{R} \to \mathbb{R}_+$, i.e., a symmetric density function which is non-increasing on $\mathbb{R}_+$. For a database $\left((x_j^1, \ldots, x_j^d, y_j)\right)_{j<t}$, consider the following formula:

$$f_t(y) = \frac{\sum_{j<t} s(x_j, x_t) K(y - y_j)}{\sum_{j<t} s(x_j, x_t)}. \tag{18}$$

This formula is an (s-)similarity-weighted average of the kernel functions $K(y_j - y)$. Thus, each observation y_j is thought of as inducing a density function $K_{y_j}(y) = K(y_j - y)$ centered around y_j. These density functions are aggregated so that the weight of $K(y_j - y)$ in the assessment of the density of y_t is proportional to the degree that the data point x_j is similar to the new data point x_t.

As in the other models discussed above, two special cases of (18) may be of interest. First, assume that s is constant. This is equivalent to suggesting that all past observations are equally relevant. In this case, (18) boils down to classical kernel estimation of the density f (ignoring the variables $x^1, \ldots, x^d$). Another special case is given by $s(x_t, x_j) = 1_{\{x_t = x_j\}}$.[7] In this case, (18) becomes a standard kernel estimation of f given only the sub-database defined by x_t. Thus, formula (18) may be viewed as offering a continuous spectrum between the unconditional kernel estimation and conditional kernel estimation given x_t.

In this section we justify the formula (18) on axiomatic grounds and develop a procedure for its estimation. We start with the axiomatic model, considering the estimated density as a function of the database. We then proceed to interpret the formula we obtain as a data-generating process. This implies that the functions s and K, whose existence follows from the axioms, can be viewed as functions of unknown parameters of a distribution, and thus as the object of statistical inference. We proceed to develop the statistical theory for the estimation of these functions.

6.1. Axiomatization

Let F be the set of continuous, Rieman-integrable density functions on $\mathbb{R}$.[8] Let $C = \mathbb{R}^{d+1}$ be the set of possible observations.[9] A *database* is a sequence of cases, $D \in C^n$ for $n \geq 1$. The set of all databases is denoted $C^* = \cup_{n\geq 1} C^n$. The concatenation of two databases, $D = (c_1, \ldots, c_n) \in C^n$ and $E = (c'_1, \ldots, c'_t) \in C^t$, is denoted by $D \circ E$ and it is defined by $D \circ E = (c_1, \ldots, c_n, c'_1, \ldots, c'_t) \in C^{n+t}$. Observe that the same element of C may appear more than once in a given database.

Fix a prediction problem, $x_t \in \mathbb{R}^d$. We suppress it from the notation through the statement of Theorem 1. For each $D \in C^*$, the predictor has a density $f(D) \in F$ reflecting her beliefs over the value of y_t in the problem under discussion. Thus, we study functions $f : C^* \to F$, and our axioms will take the form of consistency requirements imposed on such functions.

For $n \geq 1$, let Π_n be the set of all permutations on $\{1, \ldots, n\}$, i.e., all bijections $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$. For $D \in C^n$ and a permutation $\pi \in \Pi_n$, let πD be the permuted database, that is, $\pi D \in C^n$ is defined by $(\pi D)_i = D_{\pi(i)}$ for $i \leq n$.

We formulate the following axioms.

A1, Order Invariance: For every $n \geq 1$, every $D \in C^n$, and every permutation $\pi \in \Pi_n$, $f(D) = f(\pi D)$.

A2, Concatenation: For every $D, E \in C^*$, $f(D \circ E) = \lambda f(D) + (1 - \lambda) f(E)$ for some $\lambda \in (0, 1)$.

Almost identical axioms appear in BGSS (2005). They deal with probability vectors over a finite space, rather than with densities. In their model, for every database D there exists a probability vector $p(D)$ in a finite-dimensional simplex, and the axioms they impose are identical to A1 and A2 with p playing the role of f.

The Order Invariance axiom states that a permuted database will result in the same estimated density. This axiom is not too restrictive provided that the variables $x = (x^1, \ldots, x^d)$ specify all relevant information (such as the time at which the observation was made). The Concatenation axiom has the following behavioral interpretation. Assume that, given database D, an expected utility maximizer has to make decisions, where the state of the world is $y \in \mathbb{R}$, and assume that her beliefs are given by the density $f(D)$. The Concatenation axiom is equivalent to saying that, for any integrable bounded utility function, if act a has a higher expected utility than does act b given each of two disjoint databases D and E, then a will be preferred to b also given their union $D \circ E$. Equivalently, the Concatenation axiom requires that, for any two integrable bounded functions $\varphi, \psi : \mathbb{R} \to \mathbb{R}$, if the expectation of $\varphi(Y)$ is at least as large as that of $\psi(Y)$ given each of two disjoint databases D and E, then this inequality holds also given their union $D \circ E$. This axiom is a variation of the Combination axiom in Gilboa and Schmeidler (2003), where it is extensively discussed. In particular, the Combination axiom is unlikely to hold when the data may reflect patterns. Thus, when time series are involved, a straightforward application of our method may lead to poor predictions.

The following theorem is an adaptation of the main result of BGSS (2005) to our context.

Theorem 1. *Let there be given a function $f : C^* \to F$ and assume that not all $\{f(D)\}_{D \in C^*}$ are collinear. Then the following are equivalent:*

(i) *f satisfies A1 and A2;*

(ii) *There exists a function $f_0 : C \to F$, and a function $s : C \to \mathbb{R}_{++}$ such that, for every $n \geq 1$ and every $D = (c_1, \ldots, c_n) \in C^n$,*

$$f(D) = \frac{\sum_{j\leq n} s(c_j) f_0(c_j)}{\sum_{j\leq n} s(c_j)}. \tag{$*$}$$

Moreover, in this case the function f_0 is unique, and the function s is unique up to multiplication by a positive number.

Recall that the discussion has been relative to a new datapoint x_t, and that $c_j = (x_j^1, \ldots, x_j^d, y_j)$. Abusing notation, we write (x_j, y_j) for $(x_j^1, \ldots, x_j^d, y_j)$. Thus, an explicit formulation of ($*$) would be

$$f(D, x_t)(y) = \frac{\sum_{j\leq n} s\left((x_j, y_j), x_t\right) f_0((x_j, y_j))(y)}{\sum_{j\leq n} s\left((x_j, y_j), x_t\right)}. \tag{19}$$

We interpret this formula as follows. Let $s\left((x_j, y_j), x_t\right)$ be the degree to which past observation (x_j, y_j) is considered to be relevant to the present datapoint x_t. We would like to think of this degree of relevance as the similarity of the past case to the present one. Let $f_0((x_j, y_j))(y)$ be the value of the density function, given a single observation (x_j, y_j), at the point y. Then, given database D, the estimated density of y is a similarity-weighted average of the densities $f_0((x_j, y_j))(y)$ given each past observation, where more similar observations get proportionately higher weight in the average.

We now make the following additional assumptions: (i) the similarity function depends only on the variables $x = (x^1, \ldots, x^d)$,

[7] We assume that the function s is strictly positive. This simplifies the analysis as one need not deal with vanishing denominators. Yet, for the purposes of the present discussion it is useful to consider the more general case, allowing zero similarity values. This case is not axiomatized in this paper.

[8] Our results can be extended to $\mathbb{R}^m$ with no major complications.

[9] For the purposes of the axiomatization, C may be an abstract set of arbitrarily large cardinality.

thus, $s\left((x_j, y_j), x_t\right) = s\left(x_j, x_t\right)$; (ii) the density function $f_0((x_j, y_j))(y)$ does not depend on x_j, i.e., $f_0((x_j, y_j))(y) = f_0(y_j)(y)$; and (iii) the density $f_0(y_j)(y)$ is a non-increasing function of the distance between y_j and y, that is, $f_0(y_j)(y) = K(y_j - y)$ for a kernel function $K \in F$.[10] Under these assumptions, (19) boils down to (18).

We refer to (19) as a "double-kernel" density function: each observation y_j for predictor values x_j affects the density of y values that are close to y_j, and it does so not only for the density of y given the specific x_j, but also for values of x that are close to x_j.

6.2. Statistical analysis

The formula (18) can be viewed either parametrically or non-parametrically. If the former approach is taken, then (18) is assumed to be correctly specified up to a finite dimensional vector of parameters, say, $\psi = (w, \theta)'$, where $w = (w_1, \ldots, w_d)$ are the weights of the similarity function as above, and $\theta = (\theta_1, \ldots, \theta_r)$ are parameters that specify the kernel function K.[11] To estimate this model, let $\mathcal{F}_t = \sigma(x_1, \ldots, x_t, y_1, \ldots, y_{t-1})$ and assume that the true conditional density of y_t, given $\mathcal{F}_{t-1}$, is given by

$$f_t(y; \psi) = \frac{\sum_{j<t} s_w\left(x_t, x_j\right) K_\theta\left(y - y_j\right)}{\sum_{j<t} s_w\left(x_t, x_j\right)}, \quad t = 2, 3, \ldots, n.$$

The joint density of $y = (y_1, \ldots, y_n)$, conditional on $x = (x_1, \ldots, x_n)$, is

$$f(y; \psi) = \prod_{t=1}^{n} f_t(y_t; \psi) = \prod_{t=1}^{n} \frac{\sum_{j<t} s_w\left(x_t, x_j\right) K_\theta\left(y_t - y_j\right)}{\sum_{j<t} s_w\left(x_t, x_j\right)}.$$

We can proceed with any classical approach, such as maximum likelihood estimation (MLE), where the MLE of ψ is defined as

$$\hat{\psi} = \arg\max_{\psi} \sum_{t=1}^{n} \log \frac{\sum_{j<t} s_w\left(x_t, x_j\right) K_\theta\left(y_t - y_j\right)}{\sum_{j<t} s_w\left(x_t, x_j\right)}.$$

Then, the estimated conditional density of y_t is $f_t\left(y; \hat{\psi}\right)$.

Alternatively, we can take a nonparametric approach, viewing (18) as a nonparametric conditional density estimator. If we consider a kernel function given up to a single bandwidth parameter h, we obtain the following double-kernel, adaptive non-parametric density estimator,

$$f_t(y) = \frac{\sum_{j<t} s_w\left(x_t, x_j\right) K\left(\frac{y - y_j}{h}\right)}{h \sum_{j<t} s_w\left(x_t, x_j\right)} \tag{20}$$

depending on $d + 1$ parameters, $w_1, \ldots, w_d, h$. In the special case where $w_1 = \cdots = w_d = 0$ (i.e., when all the s_w's are equal), the formula reduces to the usual kernel density estimate,

$$f_t(y) = \frac{1}{(t-1)h} \sum_{j=1}^{t-1} K\left(\frac{y - y_j}{h}\right).$$

In order to make (20) operational, we can choose h and w jointly so as to satisfy any reasonable criterion, such as the minimum of the MISE.

[10] These simplifying assumptions can be written in terms of axioms on $f : C^* \to F$. However, this translation is straightforward and therefore omitted.

[11] Of course, one may consider richer parametric models, such as a quadratic distance function that depends on $\binom{d}{2} + d$ parameters.

7. Discussion

Analogical reasoning is a cornerstone of human intelligence. Formal and axiomatically based models of such reasoning have resulted in the empirical similarity approach discussed above. The formulae used in this approach turn out to be very similar to kernel methods in statistics. While the differences between the empirical similarity approach and kernel methods should not be underestimated, the striking similarity between the formulae used in both method is probably not coincidental.

Our findings suggest that a closer interaction between statistical theory and axiomatic decision theory may be fruitful for both disciplines. Statistical techniques may be interpreted as models of human reasoning and decision making. Just as kernel techniques may be viewed as formal models of reasoning by analogies, other statistical methods may also inform us regarding the way people think. In particular, regression analysis suggests a simple model of reasoning that goes beyond mere analogies to the identification of trends. It appears obvious that decision makers engage in such reasoning, and decision theory should incorporate it into its formal models.

Conversely, the axiomatic approach may further our understanding of statistical techniques and help us see connections among them. For instance, we find that a basic principle, namely the Combination axiom, appears to be at the foundation of several techniques, such as kernel estimation, kernel classification, likelihood maximization as well as the empirical similarity approach. Studying the underlying principles of various methods may suggest new ways to combine them in order to tackle new problems.

Appendix. Proof of Theorem 1

The necessity of the axioms is straightforward. We now prove sufficiency.

Consider the sequence of partitions of $\mathbb{R}$ defined by

$$P_m = \{(-\infty, -m), [m, \infty)\} \cup \left\{ \left[T + \frac{l}{2^m}, T + \frac{l+1}{2^m} \right) \,\middle|\, -m \le T \le m-1, 0 \le l \le 2^m - 1 \right\}. \tag{21}$$

Thus, P_m contains $m2^{m+1} + 2$ intervals, of which two are infinite. For $f \in F$, let f_m be the distribution induced by f on P_m. Specifically, for $A \in P_m$, $f_m(A) = \int_A f(y)\mathrm{d}y$. Observe that, for every $f \in F$, $\max\{f_m(A) \mid A \in P_m\} \to 0$ as $m \to \infty$.

Fix P_m and consider $f_m(D)$ for $D \in C^*$. Observe that f_m satisfies the axioms of BGSS (2005). Hence for every $m \ge 1$ there exists a function $s_m : C \to \mathbb{R}_{++}$ such that, for every $n \ge 1$, every $D = (c_1, \ldots, c_n) \in C^n$, and every $A \in P_m$,

$$f_m(D)(A) = \frac{\sum_{j \le n} s_m(c_j) f_m(c_j)(A)}{\sum_{j \le n} s_m(c_j)}. \tag{22}$$

It follows that (22) holds also for every event A that is P_m-measurable. Consider two consecutive partitions, P_m and P_{m+1}. Since every event $A \in P_m$ is also P_{m+1}-measurable, we conclude that, for every $n \ge 1$, every $D = (c_1, \ldots, c_n) \in C^n$, and every $A \in P_m$,

$$f_{m+1}(D)(A) = \frac{\sum_{j \le n} s_{m+1}(c_j) f_{m+1}(c_j)(A)}{\sum_{j \le n} s_{m+1}(c_j)}. \tag{23}$$

However, $f_{m+1}(D)(A) = f_m(D)(A) = \int_A f(D)(y)\mathrm{d}y$ and $f_m(c_j)(A) = f_{m+1}(c_j)(A) = \int_A f(c_j)(y)\mathrm{d}y$. Combining these with (22) and (23), we conclude that s_{m+1} can replace s_m in (22). By the uniqueness result of BGSS (2005), s_{m+1} is a multiple of s_m. Without loss of generality, we may assume that $s_{m+1} = s_m$. Thus, these

exists a function $s : C \to \mathbb{R}_{++}$, and, for each $c \in C$, a density $f(c) \in F$, such that, for every $m \geq 1$, for every $n \geq 1$, every $D = (c_1, \ldots, c_n) \in C^n$, and every $A \in P_m$,

$$f_m(D)(A) = \frac{\sum_{j \leq n} s(c_j) f(c_j)(A)}{\sum_{j \leq n} s(c_j)}. \tag{24}$$

Next consider an arbitrary finite interval (u, v) (where $-\infty \leq u < v \leq \infty$). Observe that, for every $n \geq 1$ and every $D = (c_1, \ldots, c_n) \in C^n$,

$$\begin{aligned}
f(D)((u,v)) &= \lim_{m\to\infty} \sum_{\{A \in P_m | A \subset (u,v)\}} f_m(D)(A) \\
&= \lim_{m\to\infty} \sum_{\{A \in P_m | A \subset (u,v)\}} \frac{\sum_{j \leq n} s(c_j) f(c_j)(A)}{\sum_{j \leq n} s(c_j)} \\
&= \lim_{m\to\infty} \sum_{j \leq n} \frac{s(c_j)}{\sum_{j \leq n} s(c_j)} \sum_{\{A \in P_m | A \subset (u,v)\}} f(c_j)(A) \\
&= \sum_{j \leq n} \frac{s(c_j)}{\sum_{j \leq n} s(c_j)} \lim_{m\to\infty} \sum_{\{A \in P_m | A \subset (u,v)\}} f(c_j)(A) \\
&= \sum_{j \leq n} \frac{s(c_j)}{\sum_{j \leq n} s(c_j)} f(c_j)((u,v));
\end{aligned}$$

hence $(*)$ is proved.

Finally, the uniqueness of f is obvious, and the uniqueness of s (up to multiplication by a positive number) follows from the uniqueness result in BGSS (2005). □

References

Akaike, H., 1954. An approximation to the density function. Annals of the Institute of Statistical Mathematics 6, 127–132.

Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer, Dordrecht.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society Series B 36, 192–236.

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. An axiomatization of an exponential similarity function. Mathematical Social Sciences 55, 107–115.

Cressie, N., 1993. Statistics for Spatial Data. John Wiley & Sons, New York.

Gayer, G., Gilboa, I., Lieberman, O., 2007. Rule-based and case-based reasoning in housing prices. BE Journals in Economics 7. Article 10.

Gilboa, I., Schmeidler, D., 1995. Case-based decision theory. Quarterly Journal of Economics 110, 605–639.

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

Gilboa, I., Schmeidler, D., 2003. Inductive inference: An axiomatic approach. Econometrica 71, 1–26.

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

Hume, D., 1748. Enquiry into the Human Understanding. Clarendon Press, Oxford.

Lieberman, O., 2010. Asymptotic theory for empirical similarity models. Econometric Theory 26 (in press). Mimeo.

Parzen, E., 1962. On the estimation of a probability density function and the mode. Annals of Mathematical Statistics 33, 1065–1076.

Riesbeck, C.K., Schank, R.C., 1989. Inside Case-Based Reasoning. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ.

Rosenblatt, M., 1956. Remarks on some nonparametric estimates of a density function. Annals of Mathematical Statistics 27, 832–837.

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

Scott, D.W., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley and Sons, New York.

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

Yang, L., Tchernig, R., 1999. Multivariate bandwidth selection for local linear regression. Journal of the Royal Statistical Society Series B 61 (4), 793–815.