A Similarity-Based Approach to Prediction

Itzhak Gilboa*, Offer Lieberman†, and David Schmeidler‡

November 27, 2006

Abstract

Assume we asked to assess a real-valued variable Y_p based on certain characteristics $X_p = (X_p^1, ..., X_p^m)$, and on a database consisting of $(X_i^1, ..., X_i^m, Y_i)$ for $i = 1, ..., n$. Analogical reasoning suggests to combine past observations of X and Y with the current values of X to generate an assessment of Y is *similarity-weighted averaging*. Specifically, the predicted value of Y, $\bar{Y}_p^s$, is the weighted average of all previously observed values Y_i, where the weight of Y_i, for every $i = 1, ..., n$, is the similarity between the vector $X_p^1, ..., X_p^m$, associated with Y_p, and the previously observed vector, $X_i^1, ..., X_i^m$. In a previous paper we axiomatized this rule, and suggested that the similarity function be estimated from past data. The current paper discusses this approach as a statistical method of prediction, and studies its relationship to the statistical literature.

*Tel-Aviv University, HEC, and Yale University. igilboa@tau.ac.il
†The Technion. offerl@ie.technion.ac.il
‡Tel-Aviv University and The Ohio State University. schmeid@tau.ac.il

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 fudamental tool by which we learn from the past about the future. He wrote, "In reality, all arguments from experience are founded on the similarity which we discover among natural objects... From causes which appear *similar* we expect similar effects. This is the sum of all our experimental conclusions." (Hume 1748, Section IV) Analogical reasoning has been widely studied in psychology and artificial intelligence (see...), 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 it 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, $x_t = (x_t^1, ..., x_t^d)$, and on a database consisting of the variables $(x_i^1, ..., x_i^d, y_i)$ for $i = 1, ..., n$. 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 $x_i = (x_i^1, ..., x_i^d)$ 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. Specifically, one may assume that there a similarity function $S : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{++} = (0, \infty)$ such that, given a database $(x_i, y_i)_{i \leq n}$ and a new data point $x_t = (x_t^1, ..., x_t^d) \in \mathbb{R}^d$, the estimate of y_t is

$$\overline{y}_t^S = \frac{\sum_{i \leq n} S(x_i, x_t) y_i}{\sum_{i \leq n} S(x_i, x_t)} \tag{1}$$

Observe that, in case 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 to use a weighted one, where the weights reflect the relevant similarity.

The formula (1) has been axiomatized in Gilboa, Lieberman, and Schmeidler (GLS, 2006) for the case that y is a real-valued variable, and in Billot, Gilboa, Samet, and Schmeidler (BGSS, 2005) 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 observale measure varies with the database that is supposed to be 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 interpretation of the axioms in GLS (2006), and BGSS (2005) can be descriptive or normative. As descriptive theories, they suggest that the formula (1) is a reasonable model of how people actually make assessments. Normatively interpreted, these theories can be taken as an argument for the use of (1) as a model of how people *should* make assessments. A statistical method should obviously be judged on normative grounds. Indeed, how people actually make assessments may be an empirical question of interest to psychologists, but not necessarily a theoretical question of interest to statisticians. Yet, taking an evolutionary perspective, one may be interested in studying the way the human mind makes inferences, in the hope that at least in some applications it may suggest useful ideas.

The formula above 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 to estimate the similarity function from the data. This can be done either by cross-validation, as a curve-fitting problem, or as a statistical inference problem. For the latter, one needs to assume a stochastic process where the formula (1), combined with a random factor, is assumed to generate the data. (See details in Secion 2 below.) In such a model, the estimation of the similarity function becomes a problem of statistical inference. GLS (2006) study such a statistical model, using a particular parametrized family of similarity functions, and develop statistical inference tools for that family. The term "empirical similarity"

refers to the function that is estimated from the data, either by cross validation, or by statistical inference techniques such as maximum likelihood estimation.

As formula (1) itself, the empirical similarity function can also be interpreted descriptively or normatively. From a descriptive viewpoint, one may argue that people learn how to judge similarity from their experience. For example, a child may judge similarity between cars based on their color, while an adult – based on the make and model. It is experience that tells us that, when a car's performance is of interest, the make and model are more important features of similarity judgments than is the color. Gilboa and Schmeidler (2001) refer to this process as "second-order induction", and argue that the human mind engages in such learning. From a normative point of view, it makes sense that this process, by which the similarity function is learnt from the data, can be optimized and refined to obtain better analogy-based predictions. Obviously, it is this interpretation that is the focus of this paper.

The formula (1) is mathematically equivalent to kernel estimation of a non-parametric function, where the similarity function plays the role of the kernel. Thus, the axiomatic derivations in GLS (2006) and BGSS (2005) may be viewed as axiomatizing kernel-based non-parametric methods. If one takes GLS (2006) and BGSS as descriptive models of human reasoning, one might argue that the statisical methods suggested by XXX and YYY as efficient methodologies for the estimation of a function whose functional

form is not known coincide with the way the human mind has evolved to estimate unknown 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.

Section 2 describes the empirical similarity statistical models. We devote Section 3 to a more detailed discussion of the relationship between kernel estimation and the connections between 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, and provides some empirical findings. 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

As mentioned above, finding the "best" similarity function for a given database may be viewed as a simple cross-validation problem. The cross-validation problem would take slightly different forms if the data are assumed to be ordered or not. Specifically, if there is a database $(x_i, y_i)_{i \leq n}$, where, for every $i > j$, (x_i, y_i) was realized after (x_j, y_j), it is natural to define, for a similarity function $S : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{++}$,

$$\hat{y}_i^S = \frac{\sum_{j<i} S(x_j, x_i) y_j}{\sum_{j<i} S(x_j, x_i)} \tag{2}$$

and then to ask which function S, in a pre-specified familiy, minimizes

$$SSE = \sum_{i \leq n} \left(y_i - \hat{y}_i^S\right)^2.$$

Billot, Gilboa, and Schmeidler (2005) provided 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

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

Thus, minimizing the SSE by the selection of w becomes a well-defined optimization problem with d parameters.

In the case that the order of the datapoints in $(x_i, y_i)_{i \leq n}$ is arbitrary, it

is more natural to define the estimate of y_i, for a given S, to be

$$\hat{y}_i^S = \frac{\sum_{j \neq i} S(x_j, x_i) y_j}{\sum_{j \neq i} S(x_j, x_i)} \tag{3}$$

and use these estimates for the calculation of the empirical similarity S. @As shown by Härdle and Marron (1985), the resluting estimates of S ...[NEED A DISCUSSION ON THE COMPUTATIONAL ASPECTS OF CV AND ITS ROBUSTNESS TO OUTLIERS]@

Cross validation techniques have the advantage of independence of statistical assumptions. While the empirical similarity functions computed by cross validation depend on the specific measure of the fit (such as the sum of squared errors), they do not depend on any assumptions of a statistical model. However, in order to conduct statistical inference and to obtain qualitative results by hypotheses tests, one would like to go beyond cross validation and present a complete statistical model. The most straightforward way to do that would seem to be to take the formulae (2) or (3) and use them in the data generating process, such that the expression for $\hat{y}_i^S$ becomes the expectation of y_i, with a Gaussian noise variable that is independent of past observations. Explicitly, the ordered model gives rise to the process

$$y_t = \frac{\sum_{i<t} S_w(x_i, x_t) y_i}{\sum_{i<t} S_w(x_i, x_t)} + \varepsilon_t \tag{4}$$

where $\varepsilon_t \sim N(0, \sigma^2)$, independently of past ε_i's. Similarly, the unordered

model would induce the statistical model

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

where $\varepsilon_t \sim N(0, \sigma^2)$, independently of other ε_i's.

Model (4) 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. Thus, (4) can be thought as a model of the mental process that economic agents engage in when determining prices. The estmimation 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 (2005).

Model (5) cannot be directly interpreted in the same way. Because the distribution of each y_t depends on all the other y_i's, (5) cannot be a temporal account of the evolution of the process.

Both models (4, 5) 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, which states, roughly speaking, that, should a certain conclusion be warranted given two disjoint databases, the

same conclusion should be warranted given their union. 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 (4, 5) 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 data generating process (DGP) of the following

type

$$\begin{aligned} y_i &= m(x_i) + \varepsilon_i, \qquad (i = 1, ..., n), \\ \varepsilon_i &\sim N\left(0, \sigma^2\right), \end{aligned} \tag{6}$$

where x_i is a scalar, $m : \mathbb{R} \to \mathbb{R}$ is the unknown function relating Y to X, and the noise variables ε_i are assumed to be iid. A widely-used nonparametric estimator of $m(\cdot)$ is

$$\hat{m}(x_0) = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x_t}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x_i - x_t}{h}\right)}, \tag{7}$$

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

$$\frac{1}{h} K\left(\frac{x_i - x_t}{h}\right) = \left(2\pi h^2\right)^{-1/2} \exp\left(-\frac{(x_i - x_t)^2}{2h^2}\right). \tag{8}$$

The choice of h is central in the nonparametric literature. When h is small, the kernel is "concentrated", and puts higher relative weight on close observations relative to more distant ones. This would make the estimate of $m(x_t)$ less influenced by different x values, but it would also make it largely dependent on fewer observations, and thus noisier. Increasing h adds stability to the estimation of $m(x_t)$, reducing its dependence on the few observations that happened to be gathered at close x values. But this more robust estimate will

be less precise, since it estimates a smoothed version of the function m rather than m itself. Accuracy and robustness may be traded off so as to minimine the mean integrated squared error (MISE). This leads to an optimal optimal bandwidth

$$\begin{aligned} h^* &= \arg\min_h \int E_{f_0} \left(\hat{m}\left(x\right) - m\left(x\right)\right)^2 dx \\ &= \arg\min_h E_{f_0} \int \left(\hat{m}\left(x\right) - m\left(x\right)\right)^2 dx \end{aligned} \tag{9}$$

where the expectation is taken under the true, unknown density of y.@Is this f_0? And what about the density x?@We can also write this as

$$h^* = \arg\min_h \int E_{f_0} \left(\hat{y}\left(x\right) - E_{f_0}\left(y|x\right)\right)^2 dx.$$

If x is countable, then we replace the last integral and write

$$h^* = \arg\min_h E_{f_0} \sum_i \left(\hat{y}_i\left(x_i\right) - E_{f_0}\left(y_i|x_i\right)\right)^2. \tag{10}$$

The sum is over squared deviations of the fitted values from the expected values of y. If we replace $m\left(x\right)$ by y in (9), in (10) we will have the usual sum of squared errors.

Let us compare kernel estimation to the empirical similarity approach to this problem. As described above, the empirical similarity method suggests

to estimate of y_t by

$$\hat{y}_t = \frac{\sum_{i=1}^{n} S_w(x_i, x_t)\, y_i}{\sum_{i=1}^{n} S_w(x_i, x_t)}.$$

where

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

and K is given in (8). Then,

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

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 the two models, there is a fundamental difference between them. Kernel estimation is a statistical technique that is used for the estimation of the model (6). By contrast, in models (4, 5) we

use the formula (1) as part of the data generating process itself.

This difference is accentuated when we focus on the ordered case. Model (6) assumes that the data generating process is rule-based, and that the distribution of y_t is a function of x_t alone. If the function m were known, x_t would have been a sufficient statistic for y_t. This is not the case for model (4). In this model, the data generating process is case-based, where the distribution of y_t depends on all past realizations of x, namely, x_i's for $i > t$. In (6) past observations are used to learn the general rule. But given this rule, they are immatierial. By contrast, in (4) directly affect the distribution of future variables.

Observe that this difference also has an implication regarding the type of questions that are raised about the parameter w. In (6), this parameter selects among statistical techniques. It can thus be chosen optimally, so as to minimize an expected loss function, or to have desirable asymptotic properties. But in (**??**), w may be a subject of statistical inference. Indeed, in GLS we develop tests for hypothesis of the form[1]

$$H_0 : w = 0.$$

That is, in this model "what is the true value of w?" is a meaningful question,

[1]Under the hypothesis that $w = 0$, $S_w(x_i, x_j) = 1$ 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.

whereas in (6) one may only ask, "what is a useful value of w?".

Despite these differences, the mathematical similarity between kernel-based estimation of an unknown rule and a similarity-based estimation of a case-based process may provide important insights. In particular, it is well-known that optimal estimation by kernel methods should allow the bandwidth parameter h to decrease with the size of the database. Similarly, assume that one uses the empirical similarity approach to estimate the value of y even though, in reality, the true DGP is (6). One would then expect that 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 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.[2]

One may restrict attention to the empirical similarity technique only for the estimation of DGP's that are of the form (4) or (5).[3] But if one wishes to use the empirical similarity method also when the true model is non-parametric regression, one would like to impose a condition that w grows with the size of the database.

The discussion above generalizes to higher dimensions ($d > 1$) without

[2]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.

[3]See, for instance, Gayer, Gilboa, and Lieberman, 2006, for a model of real-estate pricing where an unordered version of model (5) is assumed, with a fixed similarity function.

any fundamental modifications. Kernel estimation is used for estimation of a non-parametric model (6) where x is multi-dimensional, and the models (4, 5) have also been formulated for a multi-dimensional x. Indeed, similar relationship exist between the kernel bandwidth parameters and the weights that determine the similarity function. Specifically,

$$\begin{aligned} S_w(x_i, x_t) &= \exp(-d_w) \\ &= \exp\left(-\sum_{j=1}^{d} w_j (x_{ij} - x_{tj})^2\right) \\ &= (2\pi)^{d/2} (\det(W))^{1/2} [(2\pi)^{-d/2} (\det(W))^{-1/2} \\ &\quad \times \exp\left(-\frac{1}{2}(x_i - x_t)' W^{-1} (x_i - x_t)\right)] \\ &= (2\pi)^{d/2} (\det(W))^{1/2} \left[(\det(W))^{-1/2} K(x_i - x_t; W)\right], \quad (11) \end{aligned}$$

where W^{-1} is a diagonal matrix with elements $2w_j$, $j = 1, ..., d$, and $K(\cdot)$ is the multivariate normal density with covariance matrix W. The term in the square brackets of (11) integrates to one. In this setting

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

Finding the "empirical similarity" function reduces to finding the set of weights w that are optimal in a given sense, say, that minimize MISE. This is equivalent to finding the "best" kernel function, provided that one varies all bandwidth parameters of the kernel function (where the j'th bandwidth parameter turns out to be 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 done by XXX. 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 hypothesis of the form

$$H_0 : w_j = 0$$

suggesting that variable x_j is immaterial in similarity judgments. Rejecting such an 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 allow us to pose or test similar qualitative questions.

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) describes the two possibilities. First, the conditional density of x_i given $(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)$ is specified as

$$p_i(\cdot) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}\left\{x_i - \mu_i - \sum_{j\neq i}\beta_{i,j}\left(x_j - \mu_j\right)\right\}^2\right].$$

This results in the following joint density of the $x's$:

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

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

$$E(x_i | x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) = \mu_i + \sum_{j \neq i} \beta_{i,j} (x_j - \mu_j).$$

For example, this holds for the model

$$x_i = \mu_i + \sum_{j \neq i} \beta_{i,j} (x_j - \mu_j) + \varepsilon_i,$$

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

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

Note that B is defined as our S and is required to be non-singular. This model is also entitled *conditional autoregression* (or CAR).

These spatial models resemble models (4, 5). In fact, the latter may be viewed as special cases of the spatial model, where the similarity function S_w imposes certain restrictions on the coefficients $\beta_{i,j}$ (and where, in 4, $\beta_{i,j} = 0$ for $i < j$). The contribution of models (4, 5) can thus be viewed as suggesting a particular form of spatial models that focus on a small number

of parameters (the d weight coefficients).

@Mention also Cressie (1993), which appeared in the intro?@

5 Probability Estimation

GLS (2006) also propose to use the empirical similarity approach for the estimation of probabilities. They 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 it is the similarity-weighted average of these past observations, namely, the similarity-weighted frequency of 1's in the past:

$$\hat{p}_w\left(y_t = 1 | x_1, ..., x_{t-1}, x_t, y_1, ..., y_{t-1}\right) = \frac{\sum_{i<t} S_w\left(x_i, x_t\right) y_i}{\sum_{i<t} S_w\left(x_i, x_t\right)}. \tag{12}$$

However, there are many applications in which the given data are not ordered in any natural way. In this case, it is natural to assume that the probability that a new data point y_t equals 1 is given by

$$\hat{p}_w\left(y_t = 1 | x_1, ..., x_n, x_t, y_1, ..., y_n\right) = \frac{\sum_{i=1}^{n} S_w\left(x_i, x_t\right) y_i}{\sum_{i=1}^{n} S_w\left(x_i, x_t\right)}. \tag{13}$$

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

define

$$\hat{p}_{w,-i}\left(y_i = 1|x_1, ..., x_n, y_1, ..., y_{i-1}, y_{i+1}, ..., y_n\right) = \frac{\sum_{j\neq i} S_w\left(x_j, x_i\right) y_j}{\sum_{j\neq i} S_w\left(x_j, x_i\right)},$$

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

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

Finally, we replace (13) by

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

Note the difference between this procedure and the one discussed in Silverman (1986, pp 124-125). In our notation, Silverman's equation (6.7) reduces to

$$\hat{p}\left(y_t = 1\right) = \frac{\lambda}{n} \sum_{i=1}^{n} \left(\frac{1-\lambda}{\lambda}\right)^{(y_i - y_t)^2} \tag{14}$$

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

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

with

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

Unlike the case of nonparametric estimation of $m\left(x\right)$ with unordered data, it is not apparent how we can map λ into w. Also, with the 'right' choice of S, it is possible to find a similarity- based predicted probability which outperforms (14).

6 Double Kernel Density Estimation

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

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

$$f_t(y) = \frac{\sum_{j<t} S(x_j, x_t) K(y - y_j)}{\sum_{j<t} S(x_j, x_t)} \tag{15}$$

This formula is a (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.

Two special cases of (15) may be of interest. First, assume that S is constant. This is equivalent to suggesting that all past observations are deemed equally relevant. In this case, (15) boils down to classical kernel estimation of the density f (ignoring the variables $x = (x^1, ..., x^d)$). Another special case is given by $S(x_t, x_j) = 1_{\{x_t = x_j\}}$.[4] In this case, (15) becomes a standard kernel estimation of f given only the sub-database defined by x_t. Thus, formula (15) 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 (15) 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 unknown parameters of a distribution, and thus

[4]We assume that the function s is strictly positive. This simplifies the analysis as one need not deal with vanishing denomintaors. 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.

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 $\mathbb{R}$.[5] Let $C = \mathbb{R}^{d+1}$ be the set of possible observations. C may be an abstract set of arbitrarily large cardinality. 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, $D = (c_1, ..., c_n) \in C^n$ and $E = (c'_1, ..., c'_t) \in C^t$ is denoted by $D \circ E$ and it is defined by $D \circ E = (c_1, ..., c_n, c'_1, ..., c'_t) \in C^{n+t}$. 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 $\{1, ..., n\}$, i.e., all bijections $\pi : \{1, ..., n\} \to \{1, ..., n\}$. 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

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

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 Billot, Gilboa, Samet, and Schmeidler (2004). 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 $x = (x^1, ..., x^d)$ specify any 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 axioms is a variation of the Combination axiom in Gilboa and

Schmeidler (2003), where it is extensively discussed.

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

Theorem 1 *Let there be given a function $f : C^* \to F$. The following are equivalent:*

(i) f satisfies A1 and A2, and not all $\{f(D)\}_{D \in C^*}$ are collinear;

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

$$f(D) = \frac{\sum_{j \leq n} S(c_j) f(c_j)}{\sum_{j \leq n} S(c_j)}.$$

$*$ (16)

Moreover, in this case the function f 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 $c_j = (x_j^1, ..., x_j^d, y_j)$. Abusing notation, we write (x_j, y_j) for $(x_j^1, ..., x_j^d, y_j)$. Thus, an explicit formulation of $(*)$ would be

$$f(D, x_t)(y) = \frac{\sum_{j \leq n} S\left((x_j, y_j), x_t\right) f((x_j, y_j))(y)}{\sum_{j \leq n} S\left((x_j, y_j), x_t\right)}. \tag{17}$$

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 $f((x_j, y_j))(y)$ 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 is y is a similarity-weighted average of the densities $f((x_j, y_j))(y)$ given each past observation, where more similar observation get proportionally higher weight in the average.

We now make the following additional assumptions: (i) the similarity function depends only on the variables $x = (x^1, ..., x^d)$, thus, $S((x_j, y_j), x_t) = S(x_j, x_t)$; (ii) the density function $f((x_j, y_j))(y)$ does not depend on x_j, i.e., $f((x_j, y_j))(y) = f(y_j)(y)$; and (iii) the density $f(y_j)(y)$ is a non-increasing function of the distance between y_j and y, that is, $f(y_j)(y) = K(y_j - y)$ for a kernel function $K \in F$.[6] Under these assumptions, (16) boils down to (15).

6.2 Statistical Analysis

In this sub-section we consider a data generating process that is governed by the similarity function $S : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{++}$ and the kernel function $K \in F$. Assume that $(x_1^1, ..., x_1^d, y_1)$ has been observed, and consider $n \geq 1$. Given observations $\left((x_j^1, ..., x_j^d, y_j)\right)_{j \leq n}$, and $(x_{n+1}^1, ..., x_{n+1}^d)$, assume that y_{n+1} is distributed according to the density

$$f_{n+1}(y) = \frac{\sum_{j \leq n} S(x_{n+1}, x_j) K(y - y_j)}{\sum_{j \leq n} S(x_{n+1}, x_j)} \tag{18}$$

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

We take a parametric approach to the estimation of (17). Specifically, assume that

$$S(x_t, x_j) = e^{-d_w(x_t, x_j)}$$

where, as above,

$$d_w(x_t, x_j) = \sum_{l \leq d} w_l (x_t^l - x_j^l)^2$$

for weights $w_l > 0$.

Assume further that

$$K(\xi) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{\xi}{\sigma}\right)^2}$$

for $\sigma > 0$.

It follows that, given the first observation $(x_1^1, ..., x_1^d, y_1)$, we may write the likelihood function defined by (17), and this function will depend on $d + 1$ parameters, $(w_1, ..., w_d, \sigma)$.

7 Discussion

8 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

$$\begin{aligned}
\Pi_m &= \{(-\infty,-m),[m,\infty)\} \cup \\
&\quad \{\,[T+\frac{l}{2^m}, T+\frac{l+1}{2^m}\,| \\
-m &\leq T \leq m-1, \\
0 &\leq l \leq 2^m-1\,\}
\end{aligned}$$

Thus, Π_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 Π_m. Specifically, for $A \in \Pi_m$, $f_m(A) = \int_A f(y)dy$. Observe that, for every $f \in F$, $\max\{f_m(A) \,|\, A \in \Pi_m\} \to 0$ as $m \to \infty$.

Fix Π_m and consider $f_m(D)$ for $D \in C^*$. Observe that f_m satisfies the axioms of Billot et al. (2004). Hence for every $m \geq 1$ there exists a function $S_m : C \to \mathbb{R}_{++}$ such that, for every $n \geq 1$, every $D = (c_1, ..., c_n) \in C^n$, and every $A \in \Pi_m$,

$$f_m(D)(A) = \frac{\sum_{j\leq n} S_m(c_j) f_m(c_j)(A)}{\sum_{j\leq n} S_m(c_j)}. \tag{19}$$

It follows that (18) holds also for every event A that is Π_m-measurable. Consider two consecutive partitions, Π_m and Π_{m+1}. Since every event $A \in \Pi_m$ is also Π_{m+1}-measurable, we conclude that , for every $n \geq 1$, every

$D = (c_1, ..., c_n) \in C^n$, and every $A \in \Pi_m$,

$$f_{m+1}(D)(A) = \frac{\sum_{j \leq n} S_{m+1}(c_j) f_{m+1}(c_j)(A)}{\sum_{j \leq n} S_{m+1}(c_j)}. \tag{20}$$

However, $f_{m+1}(D)(A) = f_m(D)(A) = \int_A f(D)(y)dy$ and $f_m(c_j)(A) = f_{m+1}(c_j)(A) = \int_A f(c_j)(y)dy$. Combining these with (18) and (19) we conclude that S_{m+1} can replace S_m in (18). By the uniqueness result of Billot et al. (2004), 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 $D = (c_1, ..., c_n) \in C^n$, and every $A \in \Pi_m$,

$$f_m(D)(A) = \frac{\sum_{j \leq n} S(c_j) f(c_j)(A)}{\sum_{j \leq n} S(c_j)}. \tag{21}$$

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

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

$$= \sum_{j \leq n} \frac{S(c_j)}{\sum_{j \leq n} S(c_j)} \lim_{m \to \infty} \sum_{\{A \in \Pi_m | A \subset (u,v)\}} f(c_j)(A)$$

$$= \sum_{j \leq n} \frac{S(c_j)}{\sum_{j \leq n} S(c_j)} f(c_j)((u,v))$$

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 Billot et al. (2004). $\square$