kNN estimation of the unilateral dependency measure between random variables

Angel Cațaron  
Electronics and Computers Department  
Transylvania University of Brașov  
Romania  
Email: cataron@unitbv.ro

Răzvan Andonie  
Computer Science Department  
Central Washington University  
Ellensburg, USA  
and  
Electronics and Computers Department  
Transylvania University of Brașov  
Romania  
Email: andonie@cwu.edu

Yvonne Chueh  
Department of Mathematics  
Central Washington University  
Ellensburg, USA  
Email: chueh@cwu.edu

Abstract—The informational energy (IE) can be interpreted as a measure of average certainty. In previous work, we have introduced a non-parametric asymptotically unbiased and consistent estimator of the IE. Our method was based on the $k$th nearest neighbor (kNN) method, and it can be applied to both continuous and discrete spaces, meaning that we can use it both in classification and regression algorithms. Based on the IE, we have introduced a unilateral dependency measure between random variables. In the present paper, we show how to estimate this unilateral dependency measure from an available sample set of discrete or continuous variables, using the kNN and the naïve histogram estimators. We experimentally compare the two estimators. Then, in a real-world application, we apply the kNN and the histogram estimators to approximate the unilateral dependency between random variables which describe the temperatures of sensors placed in a refrigerating room.

I. INTRODUCTION

A critical aspect of many machine learning approaches is how well an information theory measure is estimated from the available training set. This relates to a standard concept in statistics – probability density estimation. Density estimation is the construction of an estimate of the density function from the observed data [1]. We will refer here only to nonparametric estimation, where less rigid assumptions will be made about the distribution of the observed data. Although it will be assumed that the distribution has the probability density $f$, the data will be allowed to speak for themselves in determining the estimate of $f$ more than would be the case if $f$ were constrained to fall in a given parametric family.

Much effort has been devoted to the nonparametric estimation of the mutual information (MI). The simplest approach is naïve estimation (which attempt to construct a histogram where every point is the center of a sampling interval) is plagued with both systematic (bias) and statistical errors [2]. An ideal estimator does not exist, and the choice of the estimator depends on the structure of data to be analyzed. It is not possible to design an estimator that minimizes both the bias and the variance to arbitrarily small values [2].

An alternative to the well-known MI measure is the unilateral dependency measure which can be derived from Onicescu’s IE. Onicescu’s IE was interpreted by several authors as a measure of expected commonness, a measure of average certainty, or as a measure of concentration. For a continuous random variable $X$ with probability density function $f(x)$, the IE is [3], [4]:

$$IE(X) = \int_{-\infty}^{+\infty} f^2(x)dx$$ (1)

In order to study the interaction between two random variables $X$ and $Y$, the following measure of unilateral dependency was defined by Andonie et al. [5]:

$$o(X,Y) = IE(X|Y) - IE(X)$$ (2)

with the following properties:  
1) $o$ is not a symmetric function;  
2) $o(X,Y) \geq 0$ and the equality holds iff $X$ and $Y$ are independent;  
3) $o(X,Y) \leq 1 - IE(X)$ and the equality holds iff $X$ is completely dependent on $Y$.

This measure quantifies the unilateral dependence characterizing $X$ with respect to $Y$ and corresponds to the amount of information detained by $Y$ about $X$. There is an obvious analogy between $o(X,Y)$ and the MI, since both measure the same phenomenon. However, the MI is a symmetric, not a unilateral measure.

When studying the interaction between two random variable, why is a unilateral dependency measure useful, and why do we not simply use the well-known MI? Let us consider two sets of experiments, characterized respectively by random variables $X$ and $Y$. The experiments run simultaneously and interact probabilistically. Our question is which variable influences probabilistically more the other one. Thus, $X$ can be viewed as $X|Y$ and $Y$ can be viewed as $Y|X$. While the correlation quantifies linear dependency and MI describes the degree of interdependence between two random variables, the asymmetric measure $o(X,Y)$ helps us understand which random variable, $X$ or $Y$, has a higher influence on the
other one. If both $X$ and $X|Y$ are available, we can estimate $IE(X|Y)$ as well as $o(X,Y)$, and similarly for $Y$ and $Y|X$.

When the data is acquired from real world experiments or from simulators, we need to store series of values for $X$ and $Y$ featuring the two phenomena running independently, as well as measurements of values generated by the two phenomena running simultaneously, in order to capture $X|Y$ and $Y|X$. Moreover, the precision of the $IE(X|Y)$ and $o(X,Y)$ estimators increase when more values of $X|Y$ are available for each value of $Y$.

In the present paper, we introduce two statistical inference techniques for the unilateral dependency measure $o(X,Y)$ using a) the kNN estimation method, and b) the naïve histogram estimation. Based on a simple probability distribution, we experimentally compare the two estimators. Then, in a real-world application, we apply the kNN and the kNN estimators to approximate the unilateral dependency between random variables which describe the temperatures of sensors placed in a refrigerating room with.

The rest of the paper is structured as follows. Section II gives an overview of previous work. Section III briefly describes the kNN IE approximation method. In Section IV we introduce our new kNN approximator for the $o(X,Y)$ measure. In Section V, we describe how we adapt the naïve histogram method, a standard technique, for the $o(X,Y)$ estimation. Section VI compares the kNN and histogram $o(X,Y)$ estimators. We conclude in Section VII with a synthetic comparison of the two introduced estimators.

II. PREVIOUS WORK

In a sequence of papers ([6], [7], [8], [9]), we have introduced a Parzen windows approach for the approximation of the $o(X,Y)$ dependency measure from sample datasets. We used this approach in classification and feature weighting, in combination with LVQ and Neural Gas type algorithms. The Parzen windows estimator cannot be applied on continuous spaces, and this is an important drawback. This means that the Parzen windows estimation of $o(X,Y)$ cannot be applied in an important machine learning domain – continuous function approximation (or prediction). We note that the same inconvenience exists when estimating the Shannon type MI through Parzen windows.

To overcome this problem, our first step was to introduce a kNN estimator for the IE in [10]. We proved that this estimator is asymptotically unbiased and consistent [11] (i.e., it is a “good” estimator). Let us remind ourselves that a statistic whose mathematical expectation is equal to its intended parameter is called unbiased estimator of that parameter. Otherwise, the statistic is said to be biased. A statistic that converges asymptotically to its intended parameter, as its sample size increases, is called consistent estimator of that parameter [12].

In accordance to our results from [11], we can state now that the kNN is a “good” $o(X,Y)$ estimator, both for the discrete and the continuous case.

III. $k^{th}$ NEAREST NEIGHBOR ESTIMATION OF THE INFORMATIONAL ENERGY

From our previous results [10], [11], we will summarize how we can estimate $IE(X)$ from a random sample $x_1$, $x_2$, ..., $x_n$ of $d$-dimensional observations, with a distribution having the unknown probability density $f(x)$.

The kNN estimators represent an attempt to adapt the amount of smoothing to the “local” density of data. The degree of smoothing is controlled by an integer $k$, chosen to be considerably smaller than the sample size. Let us denote $d_j(x_i)$ the distance from the reference point $x_j$ to the point $x_i$. For each $x_i$, we define $d_1(x_i) \leq d_2(x_i) \leq \ldots \leq d_n(x_i)$ to be the distances from $x_i$ to all other points of the sample, arranged in ascending order.

The kNN density estimate of probability density function $f(x)$ in $x_i$ is defined by [1]:

$$f(x_i) = \frac{k}{2nd_k(x_i)}$$ (3)

The IE is the average of the $f(x)$ values, as can be seen from eq. (1), that is $IE(X) = \bar{E}(f(x))$. Since all $n$ observations from our samples have the same probability $1/n$, a convenient estimator for IE is:

$$\hat{IE}^{(n)}_k(X) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(x_i)$$ (4)

where $\hat{f}(x_i)$ is the estimate of the probability density function $f(x_i)$.

To evaluate this formula, we have to obtain $\hat{f}(x_i)$. We start by determining the probability density $p_{ik}(\epsilon)$ of the random distance $R_{i,k,n}$ between a fixed point $x_i$ and its $k^{th}$ nearest neighbor, selected from the remaining $n - 1$ points. The probability $P_{ik}(\epsilon)de$ of the $k^{th}$ nearest neighbor, to be within distance $R_{i,k,n} \in [\epsilon, \epsilon + de]$ from $x_i$, can be expressed in terms of the trinomial formula [13]:

$$P_{ik}(\epsilon)de = \frac{(n - 1)!}{1!(k - 1)!((n - k - 1))!} dp(\epsilon) p_i^{k-1}(1-p_i)^{n-k-1}$$

where $p(\epsilon) = \int_{\|x-x_i\|<\epsilon} f(x)dx$ is the mass of the $\epsilon$-ball centered at $x_i$ and $\int P_{ik}(\epsilon)de = 1$. We notice here that the distance between $x_i$ and a subset of $k - 1$ points is smaller than $\epsilon$, while the distance to the remaining $n - k - 1$ points is larger than $\epsilon + de$.

We can express the expected value of $p_i(\epsilon)$ using the probability mass function of the trinomial distribution:

$$E_{p_{ik}(\epsilon)}(p_i(\epsilon)) = \int_0^{\infty} P_{ik}(\epsilon)p_i(\epsilon)de =$$

$$= k\left(\frac{n - 1}{k}\right) \int_0^1 p^{k-1}(1-p)^{n-k-1}dp =$$

$$= k\left(\frac{n - 1}{k}\right) \int_0^1 x^{k-1}(1-x)^{n-k-1}dx.$$
This equality can be reformulated using the Beta function:
\[ B(m, n) = \int_0^1 x^{m-1}(1-x)^{n-1} dx = \frac{\Gamma(m) \Gamma(n)}{\Gamma(m + n)}. \]

We obtain:
\[ E_{P_k(\epsilon)}(p_i(\epsilon)) = k \left( \begin{array}{c} n-1 \\ k \end{array} \right) \frac{\Gamma(k+1)\Gamma(n-k)}{\Gamma(n+1)} = k \frac{(n-1)!}{(n-k-1)!} \frac{k!(n-k-1)!}{n!}, \]
which can be rewritten as:
\[ E_{P_k(\epsilon)}(p_i(\epsilon)) = \frac{k}{n}. \] (5)

On the other hand, assuming that \( f(x) \) is almost constant in the entire \( \epsilon \)-ball around \( x_i \), we have [13]:
\[ p_i(\epsilon) \approx V_{d}R_{i,k,n}^d f(x_i) \]
where we denote the volume of the ball of radius \( \rho_{r,n} \) in a \( d \)-dimensional space by
\[ V_{\rho_{r,n}} = V_1 \rho_{r,n}^d = \frac{\pi^{\frac{d}{2}} \rho_{r,n}^d}{\Gamma(\frac{d}{2} + 1)}. \]

\( V_1 \) is the volume of the unit ball and \( R_{i,k,n} \) is the Euclidean distance between the reference point \( x_i \) and its \( k \)-th nearest neighbor. This means that \( V_1 R_{i,k,n}^d \) is the volume of the \( d \)-dimensional ball of radius \( R_{i,k,n} \). By using the Euclidean distance, we assume that all dimensions are at the same scale.

We obtain the expected value of \( p_i(\epsilon) \):
\[ E(p_i(\epsilon)) = E(V_1 R_{i,k,n}^d f(x_i)) = V_1 R_{i,k,n}^d \hat{f}(x_i). \] (6)

Equations (5) and (6) both estimate \( E(p_i(\epsilon)) \). Their results are approximatively equal:
\[ V_1 R_{i,k,n}^d \hat{f}(x_i) = \frac{k}{n}. \]
That is:
\[ \hat{f}(x_i) = \frac{k}{nV_1 R_{i,k,n}^d}, i = 1 \ldots n. \] (7)

This is the estimated probability density function. By substituting \( \hat{f}(x_i) \) in eq. (4), we finally obtain the following IE approximation:
\[ \hat{IE}_k^{(n)}(X) = \frac{1}{n} \sum_{i=1}^n \frac{k}{nV_1 R_{i,k,n}^d}. \] (8)

According to [11], \( \hat{IE}_k^{(n)}(X) \) is asymptotically unbiased and consistent (i.e., it is a “good” estimator).

IV. THE KNN \( o(X,Y) \) ESTIMATOR

Our goal is to infer \( o(X,Y) \) from the random sample \( x_1, x_2, \ldots, x_n \). We will use the results from Section III to deduct a new KNN estimator for \( o(X,Y) \).

First, we substitute \( \hat{IE}_k^{(n)}(X) \) from eq. (8) in eq. (2):
\[ \hat{o}(X,Y) = \hat{IE}_k^{(n)}(X|Y) - \hat{IE}_k^{(n)}(X) \] (9)
where:
\[ \hat{IE}_k^{(n)}(X|Y) = \sum_{j=1}^m \hat{f}(y_j) \hat{IE}_k^{(n)}(X|y_j) \] (10)
and
\[ \hat{IE}_k^{(n)}(X) = \frac{1}{n} \sum_{i=1}^n \frac{k}{nV_1 R_{i,j}^d} \] (11)
is an adaptation of eq. (8).

We can write:
\[ \hat{IE}_k^{(n)}(X|y_j) = \frac{1}{n} \sum_{i=1}^n \hat{f}(x_i|y_j) \]
where
\[ \hat{f}(x_i|y_j) = \frac{\hat{f}(x_i, y_j)}{\hat{f}(y_j)}. \] (12)

We re-write the right hand side of this equation by using the estimated probability density function from eq. (7):
\[ \hat{f}(x_i, y_j) = \frac{k_{ij}}{mnV_1 R_{i,j}^d}, i = 1 \ldots n, j = 1 \ldots m \]
\[ \hat{f}(y_j) = \frac{k_j}{mV_1 R_{j}^d}, j = 1 \ldots m \]
and we obtain:
\[ \hat{f}(x_i|y_j) = \frac{k_{ij}}{mnV_1 R_{i,j}^d} \frac{mV_1 R_{i,j}^d}{k_j} = \frac{k_{ij}R_{i,j}^d}{nk_j R_{i,j}^d}. \]

\( R_{i,j} \) is the Euclidean distance between the reference point \( x_i \) and its \( k_{ij} \)-th nearest neighbor, when the points are drawn from the one-dimensional probability distribution \( f(x) \): \( R_{i} = ||x_i - x_{ij,k_i}|| \). Similarly, \( R_{j} \) is the Euclidean distance between the reference point \( y_j \) and its \( k_{ij} \)-th nearest neighbor, when the points are drawn from the one-dimensional probability distribution \( f(Y) \): \( R_{j} = ||y_j - y_{ij,k_j}|| \). Then, \( R_{ij} \) is the Euclidean distance between the reference point \( (x_i, y_j) \) and its \( k_{ij} \)-th nearest neighbor, when the points are drawn from the joint probability distribution \( f(X,Y) \): \( R_{ij} = \sqrt{(x_{ij} - x_{ij,k_i})^2 + (y_{ij} - y_{ij,k_j})^2} \).

Now we can re-write eq. (10):
\[ \hat{IE}_k^{(n)}(X|Y) = \sum_{j=1}^m \hat{f}(y_j) \frac{1}{n} \sum_{i=1}^n \frac{k_{ij}R_{i,j}^d}{nk_j R_{i,j}^d} \]
and the estimate of \( o(X,Y) \) is:
\[ \hat{f}(x_i|y_j) = \frac{R_{ij}^d}{nR_{ij}^d}, \]

but this is not always a good option. However, although we do not have a general method to set the nearest neighbor parameter, Silverman [1] suggests that an optimal choice of \( k \) is proportional to \( n^{4/(d+4)} \).

In our case, the optimal values of \( k_i \) and \( k_{ij} \) may not be equal, because these two parameters refer to different samples.

V. HISTOGRAM ESTIMATION OF \( o(X, Y) \)

The naive histogram estimation of a probability density function is a standard technique (see [1]). We will show how we can use it for inferring \( o(X, Y) \), and this method will be an alternative to our kNN estimator from Section IV.

Considering the same random sample \( x_1, x_2, \ldots, x_n \) as above, the histogram estimator of probability density function \( f(x) \) in \( x_i \) is:

\[ \hat{f}(x_i) = \frac{\text{number of } x \text{ falling in the same bin as } x_i}{nh} \]

where \( n \) is the sample size and \( h \) is the bin width. All points falling in the same bin have the same \( \hat{f}(x) \). The empirical value of IE can be written as:

\[ IE_{\text{empirical}}(X) = \frac{\sum_{i=1}^{n} \hat{f}(x_i)}{n}. \]  

From eqs. (15) and (16), the estimate of the IE can be expressed by

\[ IE_{\text{empirical}}(X) = \frac{\text{number of bin } (n_{\text{bin}})^2}{n^2h}. \]

where \( n_{\text{bin}} \) is the number of points in the current bin. We add in each bin \( n_{\text{bin}} \) times the number \( n_{\text{bin}} \) of points falling into the same bin as \( x_i \).

In the case of the conditional probability density function, we draw for each point \( y \) several points \( x \) from \( f(x|y) \), and find \( IE_{\text{empirical}}(X|Y) \) by a formula similar to eq. (17).

The empirical value of \( o(X, Y) \), obtained from the histogram is:

\[ o_{\text{empirical}}(X, Y) = IE_{\text{empirical}}(X|Y) - IE_{\text{empirical}}(X). \]

VI. EXPERIMENTS

A. A simple probability distribution

The non-parametric estimation of the IE is appropriate when the available sample has an unknown distribution. Nevertheless, it is interesting to compare the estimator’s outcome with the results provided by a wider used technique, such as the naive histogram estimation, as well as with the theoretical value of a probability density function.

In our experiments, we consider the joint probability density function

\[ f_{X,Y}(x,y) = \frac{6}{5} (x + y^2), x \in [0, 1], y \in [0, 1], \]

which has the marginal probability density functions

\[ f_X(x) = \int_0^1 \frac{6}{5} (x + y^2) dy = \frac{6}{5} \left( x + \frac{1}{3} \right) \]

and

\[ f_Y(y) = \int_0^1 \frac{6}{5} (x + y^2) dx = \frac{6}{5} \left( \frac{1}{2} + y^2 \right). \]

The conditional probability density function is:

\[ f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{x + y^2}{\frac{1}{2} + y^2}, \]

and the theoretical value of \( o(X, Y) \) is:

\[ o(X, Y) = IE(X|Y) - IE(X) \]

\[ IE(X) = \int_0^1 f^2(x) dx = 1.12 \]

\[ IE(X|Y) = \int_0^1 \int_0^1 f_{X,Y}(x,y) f_{X|Y}(x|y) dxdy = 1.1351 \]

and

\[ o(X, Y) = 1.1351 - 1.12 = 0.0151. \]

A similar method can be used to find the values \( IE(Y) = 1.128, IE(Y|X) = 1.14787, \) and \( o(Y, X) = 0.01987. \)

To compare the theoretical values with the kNN and the histogram estimates, we need a sample of values drawn from the proposed distribution. The rejection sampling method [14] is appropriate for our case because we can find the inverses of the \( f(x) \), \( f(y) \) and \( f(x|y) \) functions.

We draw a value from \( f_X(x) \) by finding first the cumulative density function

\[ F(x) = \int_0^x \frac{6}{5} \left( z + \frac{1}{3} \right) dz = \frac{3x^2 + 2x}{5} \]

and then its inverse

\[ x = F^{-1}(t) = \sqrt{15t + 1} - 1 \]

where \( t \) is a uniform random number from \([0, 1]\).
Fig. 1. The empirical value of $o(X, Y)$ determined from the histogram, with bin width of 15 points. The theoretical values $IE(X) = 1.12$, $IE(X|Y) = 1.1351$, $o(X, Y) = 0.0151$, $IE(Y) = 1.128$, $IE(Y|X) = 1.14787$, $o(Y, X) = 0.01987$ have been marked with dashed lines.

Fig. 2. The kNN estimated value of $o(X, Y)$, where $k$ was determined with formula (14). The theoretical values $IE(X) = 1.12$, $IE(X|Y) = 1.1351$, $o(X, Y) = 0.0151$, $IE(Y) = 1.128$, $IE(Y|X) = 1.14787$, $o(Y, X) = 0.01987$ have been marked with dashed lines. $IE(X|Y)$ and $IE(Y|X)$ converge towards the theoretical values when the number of samples increase.

A sample from $f_{X|Y}(x|y)$ can be drawn with a two steps method. First, we draw a sample $y$ from $f_Y(y)$. Next, we have to determine the cumulative density function of $f_{X|Y}(x|y)$:

$$F(x|y) = \int_0^x \frac{z + y^2}{\frac{1}{2} + y^2} dz = \frac{x^2 + 2xy^2}{1 + 2y^2}$$

and

$$x_y = F^{-1}(w) = -y^2 + \sqrt{2wy^2 + y^4}$$

where $w$ is a uniform random number from $[0, 1]$ and $y$ is the sample determined at the first step.

For a better fit with the theoretical conditional probability distribution function $f_{X|Y}(x|y)$, we generate between 2–10 $x$ values for each $y$ value, and between 2000–5000 $y$ points. We apply the histogram method for 15 points in each bin because we determined experimentally that this value provided the most accurate estimations for our example. The kNN estimator is tested with the values of $k_i$ and $k_{ij}$ determined directly by formula (14), without applying any scale.

The histogram estimation (Fig. 1) looks unbiased, but has a large variability around the true values, even when the data sample size increases. Moreover, the estimated value of $o(X, Y)$ is mostly negative although it should be positive, as known from its properties. The kNN estimation (Fig. 2) is biased and tends to underestimate the true values. However, when the data sample size increases, the kNN estimation becomes more accurate, because it is asymptotically unbiased and consistent.

**B. Temperature sensors data**

In our real-world application, we study the temperature drop of two parcels placed in a room which is refrigerated by two air conditioning units $AC_1$ and $AC_2$. The experimental data are obtained with the emulator introduced in [15]. Each of the two air conditioners generate a temperature of $1^\circ$C. The two parcels $P_1$ and $P_2$ at initial temperatures $25^\circ$C and $20^\circ$C, having the sensors $TS101$ and $TS102$ attached, are placed in the room at various positions. We study the temperature variation
recorded by TS101 and TS102 under the following scenarios:

S1. Parcel P1 is placed at position POS1 to obtain the values of X measured by sensor TS101.
S2. Parcel P2 is placed at position POS2 to obtain the values of Y measured by sensor TS102.
S3. Parcel P1 is placed at position POS1. The parcel P2 is also placed, but its position slightly varies around POS2. For each new position of parcel P2, we measure a new series of values X|Y from the TS101 sensor.
S4. Parcel P2 is placed at position POS2. The parcel P1 is also placed, but its position slightly varies around POS1. For each new position of parcel P1, we measure a new series of values Y|X via the sensor TS102.

These experiments allow us to determine how the presence of a parcel in the neighborhood of the other affects the evolution of temperatures recorded by the two sensors. We
randomly re-position the parcels $P_1$ and $P_2$ five times each, denoting these scenarios by $S3.1$–$S3.5$ and $S4.1$–$S4.5$. The emulated scenarios $S1$, $S2$, $S3.1$, and $S4.1$ are presented in Fig. 3. The temperatures recorded in the above mentioned scenarios after 5000 ticks are depicted in Fig. 4. From the simulation of scenarios $S1$ and $S2$, we obtain the samples of random variables $X$ and $Y$. For each value of $X$, we obtain one corresponding value of $X|Y$ from scenario $S3.1$. For additional precision, for each value of $X$, we can obtain a set of values of $X|Y$ if we run the scenarios $S3.1$ – $S3.5$. The same idea applies for $Y$ and $Y|X$.

Table I summarizes the experimental results obtained with the histogram estimator. The histogram method yields nearly constant values for $IE(X|Y)$ and $IE(Y|X)$, given a fixed number of bins, meaning that increasing the data set has little impact on this estimator. Nevertheless, the values of $o(X, Y)$ and $o(Y, X)$ should be positive, thus the bias is an important element in this case.

Table II summarizes the experimental results obtained with the kNN estimator. The kNN estimator of the IE becomes more relevant if this measure is incorporated into machine learning algorithms. As we did in [6], [7], [8], and [9], in classification algorithms, all three estimators (Parzen windows, kNN, and histogram) may be used to approximate $o(X, Y)$ from data samples. In the future, it may be interesting to compare their relative performances. For continuous function approximation, only the kNN and the histogram methods may be used.

VII. CONCLUSIONS

In our examples, summarized in Figs. 1 and 2, all with unidimensional variables, histogram estimation is computationally more efficient than the kNN method and is more stable than the kNN estimator (smaller variance and smaller bias). However, for data sets with two or more dimensions, the histogram method becomes computationally expensive, due to the rapid increase of the number of bins. Another known drawback of the histogram method is the bias generated by the origin and width of the bins have a strong influence. It is difficult to find optimal values for the parameters of the bins (i.e., their hyper-volume, origin, and orientation) [1].

It is true, the kNN estimator is computationally more intensive than the histogram method (at least for unidimensional variables), but it is a “good” estimator – it is asymptotically unbiased and consistent – which are nice properties that the histogram method lacks. Therefore, it is generally more accurate than the histogram estimator.

The histogram method may be used, for instance, to process in real time a data stream, and draw a quick (but not final) conclusion. The kNN method is perhaps more useful in batch mode, as a second step, to consolidate the results.

Approximating $o(X, Y)$ from sample datasets becomes more relevant if this measure is incorporated into machine learning algorithms, as we did in [6], [7], [8], and [9]. In classification algorithms, all three estimators (Parzen windows, kNN, and histogram) may be used to approximate $o(X, Y)$ from data samples. In the future, it may be interesting to compare their relative performances. For continuous function approximation, only the kNN and the histogram methods may be used.

REFERENCES


