International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 7 Issue: 9 16 - 23

An Algorithm for Computing Average Mutual Information using Probability Distribution Smoothing

Amr Goneid
Department of Computer Science & Engineering,
The American University in Cairo,Cairo, Egypt
e-mail: goneid@aucegypt.edu

Abstract **—**There is continuing interest in using Average Mutual Information (AMI) to quantify the pair-wise distance between dataset profiles. Among several algorithms used to find a numerical estimation of AMI, the histogram method is the most common since it provides simplicity and least cost. However, this algorithm is known to underestimate the computed entropies and to overestimate the resulting AMI. Kernel Density Estimator (KDE)-based algorithms advanced to alleviate such systematic errors rely on bin-level smoothing. In the present work, we propose an alternative algorithm that uses smoothing on the probability distribution level. We consider several smoothing functions, both in the probability space and in its frequency space. An experimental approach is used to investigate the effect of such modification on the computation of both the entropy and the AMI. Results show that, to a significant extent, the present method is able to remove systematic errors in computing entropy and AMI. It is also shown that the present algorithm leads to better reconstruction of multivariate time series when AMI is used in conjunction with their independent components.

Keywords: *Average Mutual Information estimation, Entropy computation, Time Series Reconstruction*

*****

I. INTRODUCTION

Currently, there is a growing interest in the methodologies for clustering multivariate datasets. In most clustering algorithms, reliance is mainly on some distance measure to quantify the pair-wise distance between dataset profiles. Usually, the efficiency of the clustering process depends not only on the clustering algorithm but also on the chosen distance measure [1].It is now recognized that the framework of information theory [2] can provide a more *general* measure of dependencies between variables in contrast to linear correlation, which utilizes only up to the second order statistics. In particular, the use of Mutual Information, or rather more common the Average Mutual Information (AMI), as a measure of distance between variables, is becoming increasingly popular [3, 4]. This is because AMI can provide a better and more general measure of dependencies between datasets. Examples of using AMI to measure such dependencies are given in [5, 6, 7].

There exist several algorithms for computing a numerical estimation of the AMI [4]. As will be shown later, the most straightforward (and therefore the least cost) method is the widely used histogram method. This method is well known to *underestimate* the computed entropies due to systematic errors resulting from finite small sized samples of the dataset values The resulting effect on the computed AMI is that it is systematically *overestimated* [3].

Other algorithms have been introduced to remove systematic errors characteristic of the histogram method. Of such algorithms, we mention those using adaptive partitioning [8] and the Kernel Density Estimator (KDE) algorithms [9]. Of particular interest is the KDE method that relies on smoothing rectangular bins by the use of a generalized weight or kernel function, usually a Gaussian one. Therefore, in effect the KDE is a smoothing method *on the bin level*.

In the present work, we consider an alternative algorithm for computing the entropy and AMI that is based on smoothing *on the probability distribution level* rather than on the bin level. Smoothing is applied to the Probability Mass Function (PMF) $P(x)$ of a variable x, as well as the joint probability $P(x,y)$ for two variables x and y resulting from the rectangular binning process. We consider several smoothing functions, both in the probability space and in its frequency space. An experimental approach is used to investigate the effect of such modification on the computation of both the entropy of a dataset and the AMI between two datasets.

The present paper is organized as follows: Section II introduces the concept of AMI and gives present experimental results of its estimation based on the histogram method. Section III introduces our methods for estimating AMI using probability distribution smoothing. Sections IV and V give experimental results on artificial datasets and the reconstruction of actual financial datasets, respectively. Finally, Section VI presents the summary and conclusion of our work.

II. ESTIMATION OF AMI FROM FINITE DATASETS

A. Entropy and Average Mutual Information

Consider a dataset of size N for a random variable $X = \{x_1, x_2, .., x_N\}$ to be partitioned into M possible states. For such a variable, the Shannon entropy is defined as:

$$H(X) = -\sum_{x \in X} P(x) \log P(x) \quad (1)$$

where $P(x)$ is the probability distribution of X. On the other hand, for two such datasets X, Y (for simplicity having equal sizes N), Mutual Information (MI) measures the information that X and Y share, i.e., it measures how much knowing one of these variables reduces uncertainty about the other. In this sense, mutual information is the average amount of decrease

of uncertainty of X by observing Y, i.e., it is the average information that Y gives about X. MI is defined through the Kullback-Leibler divergence D of the joint probability distribution to the product of its marginals:

$$I(X,Y) = D(P(x,y)||P(x)P(y)) = \sum_{x\in X}\sum_{y\in Y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)} \quad (2)$$

where $P(x,y)$ is the joint probability distribution of X and Y and $P(x)$ and $P(y)$ are the marginal probability distributions. We notice here that the MI is in fact the *Expectation* relative to $P(x,y)$ of the logarithm of $P(x,y) / [(P(x) P(y)]$. This indicates that $I(X,Y)$ is a result of an averaging process and so we may also call it the *Average Mutual Information* (AMI). Moreover, the logarithm can have an arbitrary base and we will always use it to indicate *log* to the base *2*.

In terms of the Shannon entropies, AMIcan also be expressed as:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \quad (3)$$

where $H(X,Y)$ is the joint entropy and $H(X)$ and $H(Y)$ are the marginal ones.

We mention here two important properties of AMI; the first is that it is symmetrical, i.e., $I(X,Y) = I(Y,X)$, and the second is that $I(X,Y) = 0$ iff X, Y are independent.

B. Histogram Method for Computing Entropy and AMI

The most straightforward method for computing entropies H and AMI is to approximate the probabilities $P(x)$, $P(y)$ and $P(x,y)$ using histograms. In this method, the dataset points are allocated to M fixed width bins. As a preprocessing step, it is usual to remove the effect of origin point by subtracting $x_{min}$from all values of x in a dataset X. A normalization process may also be included so that $x_i \in [0,1]\ \forall i = 1,..,N$, where N is the size of the dataset.As for the optimal number of bins, it has been recommended [10] to use $M = M_{opt} = N_{bins} = \lfloor (1+\log_2(N)+0.5) \rfloor$.

In the 1-D case with $x_i \in [0,1]$ and M bins in the histogram, the fixed width of a bin will be $\Delta x = 1/M$ so that bin centers will be located at:

$$z_k = (2k-1) / (2M),\ k = 1..M \quad (4)$$

In this case, the frequency of values of X falling in a bin with center at z_k will be:

$$n(z_k) = \sum_{i=1}^{N} \Theta\left(\frac{1}{2M} - |z_k - x_i|\right) \quad (5)$$

where $\Theta(r)$ is the Heaviside function given by:

$$\Theta(r) = \begin{cases} 1 & r > 0 \\ 0 & r \le 0 \end{cases} \quad (6)$$

Computationally, we can assign for each value x_i a bin index $u_i = \lfloor M\ x_i \rfloor + 1$so that $1 \le u_i \le M$. The frequency of Xvalues allocated to a bin k will then be:

$$n(k) = \sum_{i=1}^{N} \delta(k,u_i), \qquad k = 1..M \quad (7)$$

where $\delta(i,j)$ is the Kronecker delta function given by:

$$\delta(i,j) = \begin{cases} 1 & if\ i = j \\ 0 & if\ i \ne j \end{cases} \quad (8)$$

The probability distribution $P(x)$ can then be obtained by normalization such that:

$$\sum_{k=1}^{M} n(k) = 1.0 \quad (9)$$

Similarly, in the 2-D case with $x_i \in [0,1]$ and $y_i \in [0,1]$, and M xM bins in the histogram, a pair (x_i , y_i) will be assigned bin indices $u_i = \lfloor M\ x_i \rfloor + 1$, *and* $v_i = \lfloor M\ y_i \rfloor + 1$ so that $1 \le u_i \le M$ and $1 \le v_i \le M$. In this case, the frequency of values of X,Y allocated to a bin (k,j) will be:

$$n(k,j) = \sum_{i=1}^{N} \delta(k,u_i)\delta(j,v_i), \qquad k,j = 1..M \quad (10)$$

As in the 1-D case, the joint probability distribution $P(x,y)$ is obtained by normalization of $n(k,j)$.The marginal probability distributions $P(x)$ and $P(y)$ can then be obtained by summing on the rows and columns of $P(x,y)$.

It should be noticed that the complexity of this straightforward histogram method is $O(N + M)$ for computing entropy using equation (1) and $O(N + M^2)$ for computing AMI using equation (2).

It is known that the histogram method is affected by systematic errors that tend to underestimate the entropies computed for finite datasets [3], particularly when their sizes are small. The entropy observed from the histograms is considered to be approximately given by:

$$\langle H_{obs} \rangle \approx H - (M-1) / (2N) \quad (11)$$

where H is the correct entropy [3]. This approximation is supposed to be independent of the probability distribution concerned. It is also shown [3] that these systematic errors will result in an overestimation of the observed AMI such that:

$$\langle I_{obs}(X,Y) \rangle \approx I(X,Y) + (M^2 - 2M + 1) / (2N) \quad (12)$$

Here, $I(X,Y)$ is the correct AMI and it is assumed that in the binning process, the same number of bins M is used for both datasets X and Y.

C. Experimental Results for Entropy using the Histogram Method

In order to verify the effects of the above mentioned systematic errors, we have conducted experiments to compute entropies H and AMI estimates using the straightforward histogram method. In the experiments for computing entropies, we used random numbers generated from a uniform distribution with dataset sizes N = {100,200,300,500,700, 1000, 2000 and 5000}. The number of bins allocated for each size is given by $M = \lfloor (1+\log_2(N)+0.5) \rfloor$. Since the dataset values are binned into M finite states with supposedly equal probabilities, the theoretical entropy corresponding to a given M is $H(theor)$

$= log_2 M$. For each of the given dataset sizes, the experiment was repeated $N_{exp} = 1000$ times.

Fig. 1 shows the results obtained for the difference between *H(theor)* and the observed entropy as a function of $log_{10}(N)$.It can be seen from the figure that the systematic errors in entropy computing actually decrease with dataset size *N*.

Fig. 2 shows the difference as a function of the quantity *(M-1)/(2 N)*.It can also be seen from Fig. 2 that the error in observed entropy exhibits a linear behavior with the increase in the ratio *(M-1) / (2N)* as given by the approximation (11). Actually our results give:

$$\langle H_{obs} \rangle = H - c(M-1)/(2N), \quad c = 1.4738 \qquad (13)$$

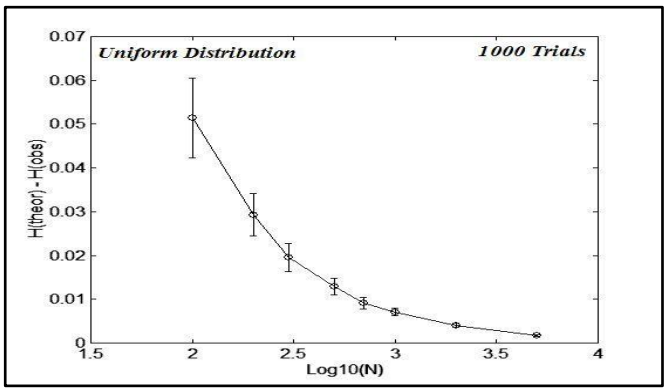

Figure 1. Entropy error as a function of dataset size.

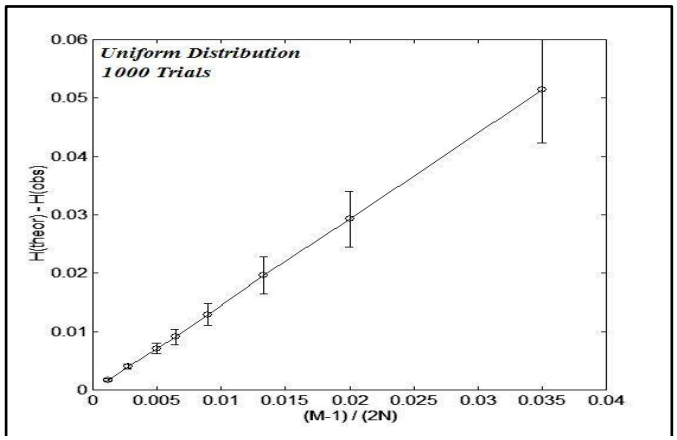

Figure 2. Entropy error as a function of (M-1)/(2N).

D. Experimental Results for AMI using the Histogram Method

To illustrate the impact of systematic binning errors on AMI, it is well known [e.g. 11] that if (*X*,*Y*) are bivariate normal, then the AMI between *X* and *Y* depends only on the correlation coefficient ρ between them.Specifically in this case, the theoretical AMI will be given by:

$$I(X,Y) = -(1/2)\, log_2(1-\rho^2) \qquad (14)$$

We have generated bivariate random datasets from a standard normal distribution with different correlations ρ. Fig. 3 shows the results for the observed AMI obtained for sizes $N = 100$ and $N = 5000$ as compared with the theoretical one (equation(14))for different correlation values ρ. It can be seen from this figure that the errors in AMI are quite significant for small dataset sizes where the AMI is significantly overestimated. However, for large values dataset sizes*N*, the errors become acceptably small over the whole range of correlation values.

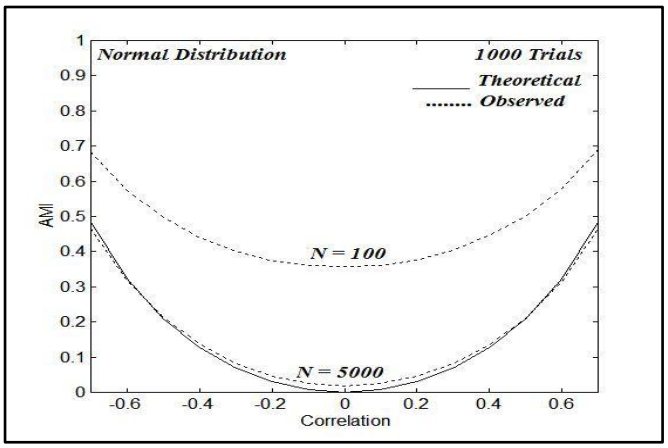

Figure 3. AMI as a function of correlation coefficient

E. Other Algorithms for Computing Entropy and AMI

In the literature, there exist several algorithms for computing entropy and AMI that are more efficient than the straightforward histogram method. Significant among these are
algorithms that use adaptive partitioning and the Kernel Density Estimator (KDE) algorithms. As an example of the adaptive partitioning algorithms, we mention that in [8]. In such algorithms, the bin widths for the histogram classes are selected by an adaptive method. This produces a partition of the data to roughly balance the proportions in the different classes.

For the KDE algorithms, the approach aims at improving the estimate of the probability density *P(x)*, and to be able to specify more sophisticated window shapes than the rectangular window for frequency counting [e.g. 12]. We may note here that the frequency distribution given by equation (5) uses the Heaviside function $\Theta(r)$ to produce rectangular windows. Such function can be replaced by some generalized kernel function *K(r)* that can provide smoothing of the bin shape. We mention here the algorithm in[12] that uses a kernel density estimator to estimate AMI. The kernel is a Gaussian that is standard normal in both the univariate and bivariate cases. In the 1-D case, a *smoothed* frequency count at a value z will be given by the estimator:

$$n(z) = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{N} exp\left(-\frac{(z-x_i)^2}{2h^2}\right) \qquad (15)$$

where *h* is the window width.

In general, the KDE method relies on smoothing rectangular bins by the use of a generalized weight or kernel

function. Therefore, in effect, the KDE is a smoothing method *on the bin level*. More details on this method can be found in [9].

III. COMPUTING ENTROPY AND AMI USING PROBABILITY DISTRIBUTION SMOOTHING

In the present work, we consider an alternative approach for computing the entropy and AMI that is based on smoothing *on the probability distribution level* rather than on the bin level. This method is applied to the Probability Mass Function (PMF) $P(x)$ of a variable X, as well as the joint probability $P(x,y)$ for two variables X and Y resulting from the rectangular binning process. While there exist many smoothing methods to apply, we have chosen to adopt three smoothing methods that operate in the probability space or in its frequency space. These methods are described as follows:

A. The Med Method:

This smoothing method is applied to the probability space and uses a 3-point moving median filter in the 1-D case for smoothing $P(x)$ and a 3 x 3 median filter in the 2-D case for smoothing the joint probability $P(x,y)$. In practice, the 2-D space of $P(x,y)$ may contain a number of zeros. The advantage of this smoothing method is that it can retain the essential features of the probability distributions in the presence of zeroes in that probability space. It is also very efficient as its complexity is only $O(M)$ in the 1-D case and $O(M^2)$ in the 2-D case, where M is the number of bins used.

B. The DCT Method:

This smoothing method operates on the frequency space of the probability distributions by using a Discrete Cosine Transform (DCT). In this method, the probability distribution with M bins is used to obtain M DCT coefficients A_i , $i = 1,..,M$. The absolute values of these coefficients are sorted in descending order and a subset of $k < M$ coefficients is selected such that their total energy does not exceed a certain percentage of the total energy in the M coefficients. The remaining $M - k$ coefficients are removed (set to zero). The smoothed distribution is then obtained by using an inverse DCT. In the 2-D case, the process is done first along the rows then along the columns and an average is obtained to represent the final smoothed distribution.

The DCT method is also very efficient as its complexity in the 1-D case is only $O(M \log M)$ and $O(M^2 \log M)$ in the 2-D case.

C. The MP Method:

In this "*Moment-Preserving*" method, smoothing the probability distribution is achieved by deriving an approximation in the probability domain while preserving a finite number of geometric moments that are related to its Fourier domain. Moment-preserving values of the probability distribution are obtained at specific nodal points that are then joined to produce the final smoothed distribution by some high order interpolation method such as quadratic or spline interpolation. Details of this MP algorithm are given in our previous work [13].

Although the MP method generally gives good smoothing results [13], it has the highest complexity compared to the other two methods mentioned above. This is because it has to compute probability values at nodal points followed by an interpolation process.

IV. EXPERIMENTAL RESULTS FOR ARTIFICIAL DATASETS

A. Entropy Estimation

We have conducted experiments to compute entropies using the histogram method with and without probability density smoothing. As an example of the effect of smoothing, Fig. 4 shows an example of the results obtained for datasets artificially generated from a uniform distribution when the DCT method is used for smoothing.

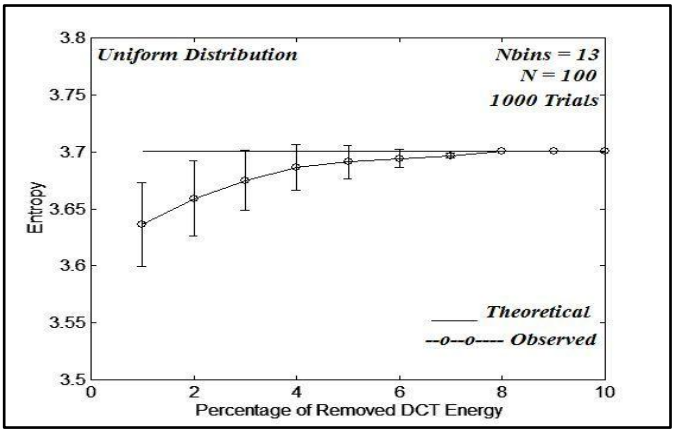

Figure 4. Effect of DCT smoothing on entropy.

The example shown uses a small dataset size $N = 100$ and a number of bins $M = 13$. The entropy error is observed to decrease by increasing the fraction of removed DCT energy and removal of only 6 – 8% of such energy diminishes the errors resulting from the histogram binning.

Fig.5 shows the ratio of observed entropy to the theoretical one for a uniform distribution as a function of dataset size$N$. The figure also shows a comparison of the smoothing effects resulting from using the three different smoothing methods described earlier.

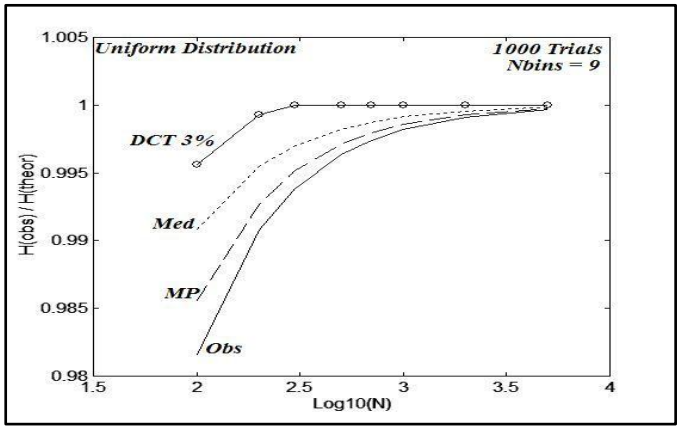

Figure 5. Comparison of smoothing methods for the entropy of a uniform distribution.

From Fig. 5, it can be seen that the DCT method gives the best results for the correction of entropy resulting from histogram binning. The Med method also gives acceptable

results but the MP method is not significantly effective. The MP method also suffers from limitations of the chosen bin sizes andfrom the relatively high complexity due to the computations of values of the probability distribution at the nodal points. Due to such limitations of the MP method, we have limited the experimentation toinclude only the Med and DCT methods.

B. Estimation of AMI

We have also conducted experiments to compute the AMI using the histogram method with and without probability density smoothing. In these experiments, we have used artificial independent datasets generated from uniform or normal distributions. For such independent datasets, the theoretical value of AMI should be zero. As an example of the effect of smoothing using the DCT method, Fig. 6 shows the smoothing effect on the error in computing AMI using a medium size datasets (N = 500) and M = 10 bins.

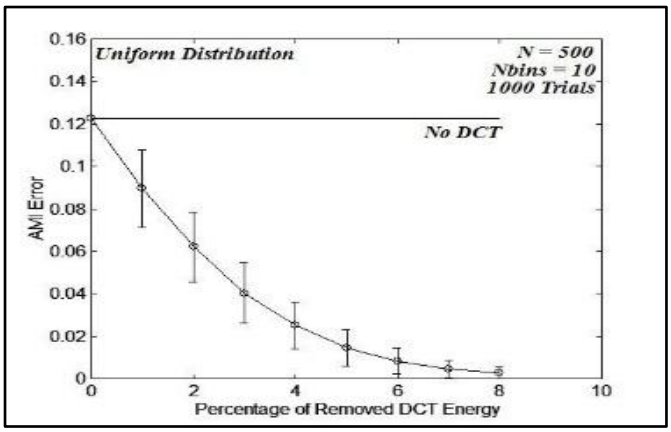

Figure 6. Effect of DCT smoothing on the error in AMI.

In Fig. 6, the "No DCT" line indicates the error in AMI when no smoothing is used, and as shown, the AMI is overestimated in this case. The results shown also indicate that the error significantly decreases with increasing the percentage of removed DCT energy. Moreover, the error in AMI diminishes to an acceptable degree by removing only between 6% and 8% of the DCT energy.

We have conducted further experiments using independent datasets to investigate the dependence of the error in AMI on the dataset size. In these experiments, the AMI error is computed with smoothing the probability densities using both the DCT and Med methods. Fig. 7a gives the observed error in AMI as a function of the dataset size $N$for independent uniformly distributed datasets. The number of bins allocated for each size is taken as$M = \lfloor (1+log_2(N)+0.5) \rfloor$. Fig. 7b shows similar results but for independent normally distributed datasets. In both Fig.7a and Fig. 7b, a comparison is made between the effects of using DCT and the Med smoothing methods on the AMI error.

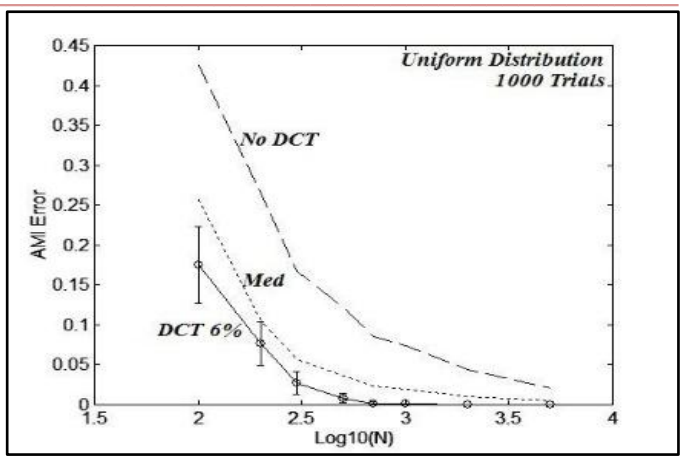

Figure 7a. Effect of DCT and MP smoothing on the error in AMI for Uniform distributions.

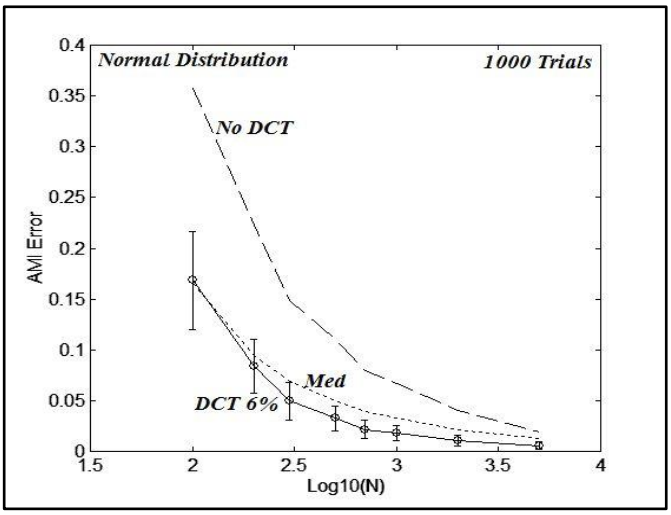

Figure 7b. Effect of DCT and MP smoothing on the error in AMI for Normal distributions

The results obtained show that, although the AMI error generally decreases by increase of dataset size, probability density smoothing can remove a significant part of the error, particularly for small and medium dataset sizes. The results also indicate that the effects of DCT and Med smoothing methods are close in decreasing the error in the computed AMI. This is particularly evident in the case of independent normally distributed datasets.

In other experiments to show the effect of DCT and Med smoothing methods, we have generated bivariate random datasets (X,Y) with standard normal distributions and characterized by different values of correlation ρ. For such datasets, the theoretical AMI will be as given before by equation (14).Fig. 8 shows the results for the observed AMI obtained for dataset size N = 500 with and without smoothing as compared with the theoretical one for different correlation values ρ.

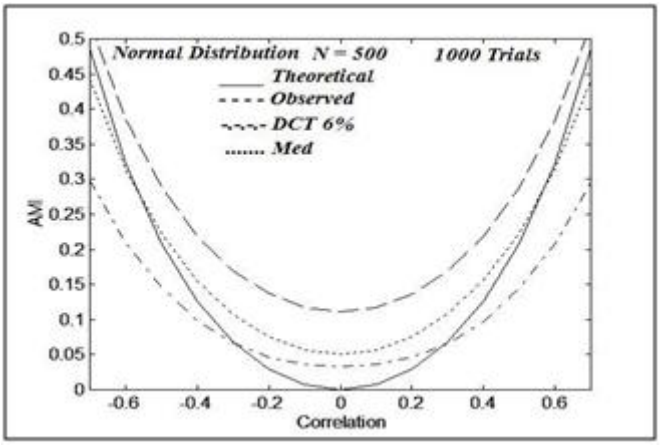

Figure 8. AMI as a function of correlation coefficient.

It can be seenfrom Fig. 8 that probability density smoothing significantly reduces the AMI errors resulting from histogram binning. Although the DCT method seems to be better in removing the error for small correlation values, we can also observe that it departs from the theoretical values at medium and high correlations. The reason seems to indicate that removal of some of the DCT energy will reduce the correlation effects in the joint probability distribution. On the other hand, we can observe that the Med method significantly reduces the errors in AMI over the greater part of the range of correlations.Preferred features in the Med method are that it has lower complexity and that it does not remove genuine correlations over most of the correlation values.

V. AN APPLICATION: RECONSTRUCTION OF REAL TIME SERIES FROM INDEPENDENT COMPONENTS

As an application of the present methodology for computing AMI, we consider the problem of reconstructing real multivariate time series from their dominant Independent Components (IC's). For this purpose, we have selected 4 financial series representing the daily exchange rates of USD versus Canadian Dollar (CAD), Euro (EUR), Pound Sterling (GBP), and Japanese Yen (JPY).The data were collected from [14] and represents a size of N = 1148 days in the period from January 2, 2015 until July 31, 2019. Fig. 9a shows these 4 financial time series over the indicated time period.

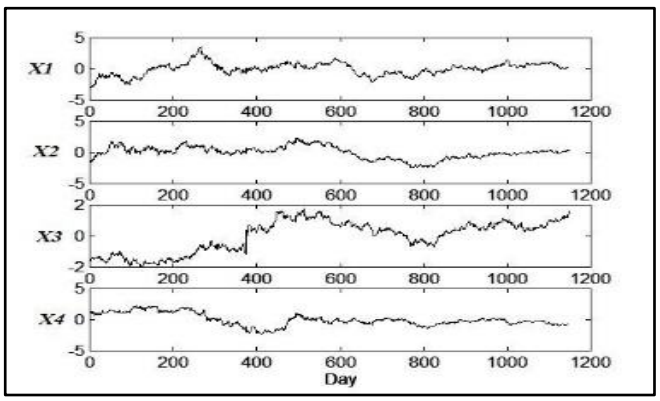

Figure 9a. Exchange Rate time series *X*.(USD versus *X1*: CAD, *X2*: EUR, *X3*: GBP, *X4*: JPY).

From the given series considered as a mixture $X = \{xi, i = 1..4\}$, we have obtained the corresponding IC's $Y = \{Yj, j = 1..4\}$ and the demixing matrix W such that an estimate of the original series is obtained as $X^* = W^{-1} Y$. The ICA algorithm used for obtaining Y and W is described in detail in our previous work [15].

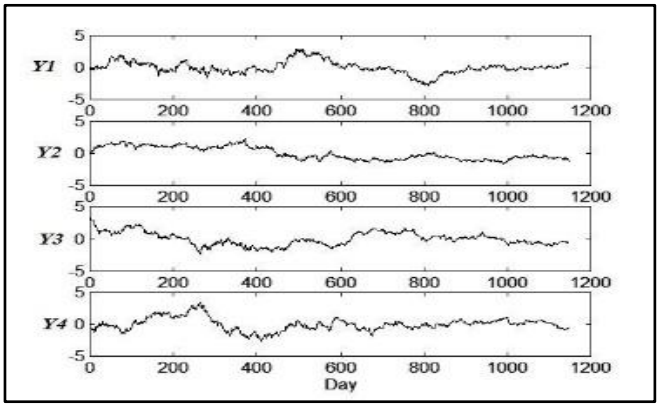

Figure 9b. IC's *Y* for Exchange Rate time series *X*.

Basically, the algorithm uses a fixed-point iteration method to maximize the negentropy using a Newton iteration method as well as a *tanh* non-linearity. The results obtained for the IC's *Y* are shown in Fig. 9b.

The process of reconstructing time series x_i from the estimated independent components Y_j, $j = 1...k$ can be done by summing their contributions in the order given by an optimal list L_i. Such list represents the indices in descending order of their dominance of contribution to a given series x_i. Following [15], the contribution may be expressed by the 3-D space:

$$u(i,j,t) = W^{-1}(i,j)\, Y_j\,(t) \quad (16)$$

The reconstruction of series of x_i by the first $m$most dominant independent components in the list L_i is obtained by summing the contributions of the individual component, i.e.

$$\hat{x}_{L_i}{}^m = \hat{x}_i^m(L_i,t) = \sum_{s=1}^{m} u(i,s,t) \quad (17)$$

where (s) denotes the s^{th} index in list L_i.

To determine the optimal list $L_i$for a given series x_i, we have computed the ordered set of IC indices that maximizes the AMI between the contributions $u(i,j,t)$ and the series x_i. We have computed the AMI's without (Algorithm A1) and with probability distribution smoothing (Algorithm A2) using the Med method outlined previously. As examples, Table (1) gives the results obtained for the USD/GBP and USD/JPY series. In the table, the optimal lists are given as obtained by algorithms A1 and A2 together with the percentage cumulative contributions of the IC's from the lists to the reconstruction of the exchange rate time series.

TABLE 1. ORDERED LISTS (*L*) AND CUMULATIVE CONTRIBUTIONS (*CC*) OF IC'S (%)(X3: USD/GBP, X4: USD/JPY)

X	X3				X4			
L (A1)	1	2	3	4	2	4	3	1
CC	5.9	69.3	91.2	100	12.9	53.7	84.6	100
L (A2)	2	1	3	4	4	2	3	1
CC	63.4	69.3	91.2	100	40.9	53.7	84.6	100

It can be seen from the above table that using AMI with smoothing has changed the order of the first two dominant IC's in the lists leading to a significantly better reconstruction using the most dominant IC. In particular, using only the first dominant IC with Algorithm A1 leads to reconstruction Mean Square Errors (MSE) of 0.94 and 0.87 for the series X3 and X4, respectively.When probability density smoothing is used (Algorithm A2), the corresponding reconstruction MSE are reduced to only 0.37 and 0.59, respectively.

This is also illustrated in Fig.10, whichcompares between a part of the observed series X3: USD/GBP and the series reconstructed from the most dominant IC in the lists obtained by the two algorithms A1 and A2. It is clear from such comparison that using probability smoothing in computing the AMI will reduce the reconstruction errors and leads to a better contribution of the dominant IC to the reconstructed series.

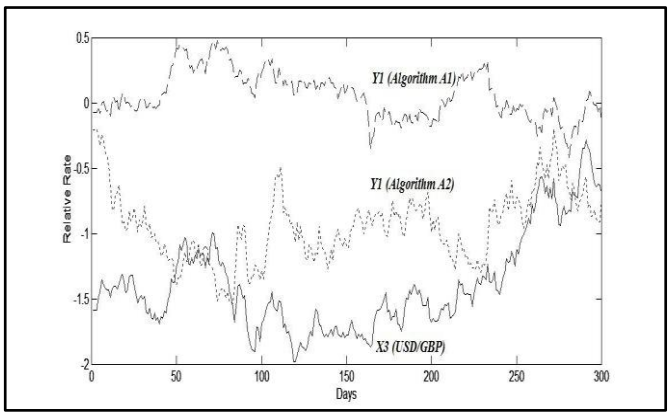

Figure 10. Observed X3 series and contributions of most dominant IC

It should be noted that exact agreement (MSE ≈ 0) between observed and reconstructed series is obtained by using all IC's in the list *L* as shown in Fig. 11. The figure compares the observed USA/GBP series with the reconstructed ones using the first one, the first two, first 3 and all IC's in the ordered list determined by AMI computations using Algorithm A2.

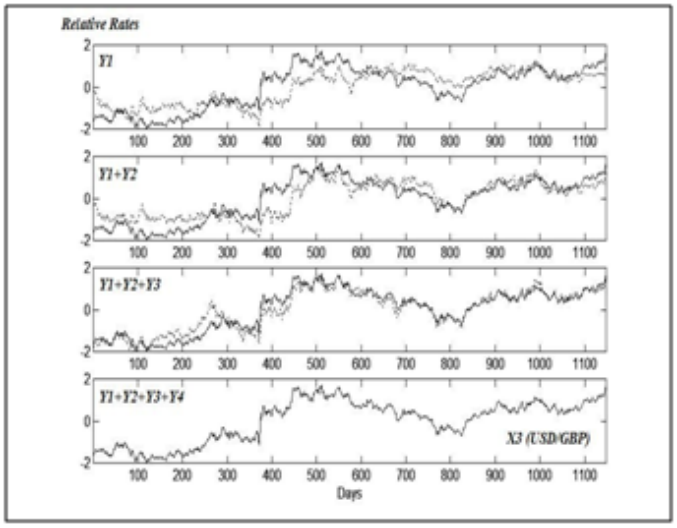

Figure 11. Reconstruction of exchange rate time series *X3 (*USD/GBP), Solid line (Observed), dotted line (Reconstructed).

It can be seen from Table (1) and Fig. 11 that the reconstruction of observed series with one or two dominant IC's can preserve the general trends of the series. Quite acceptable matching can be realized with only the first dominant 3 IC's in the lists (e.g. their contribution to the X3 series is 91.2%). Of course, exact matching is achieved when all IC's are used in the reconstruction process.

VI. SUMMARY AND CONCLUSIONS

The histogram method is considered to be the most straightforward method for computing entropies and Average Mutual Information (AMI) between datasets. It is known that the histogram method is affected by systematic errors resulting from binning dataset variables using rectangular fixed width bins. Kernel Density Estimator (KDE)-based algorithms advanced to alleviate such systematic errors rely on bin-level smoothing. In the present work, we have introduced an alternative algorithm that uses smoothing on the probability distribution level. We considered several smoothing functions, both in the probability space and in its frequency space. An experimental approach is used to investigate the effect of such modification on the computation of both the entropy and the AMI.

In order to quantify the effects of the systematic errors resulting from the fixed bin width binning, we have conducted a set of experiments to compute entropies and AMI estimates using the straightforward histogram method. Using artificial datasets of different sizes generated from uniform and normal distributions, we obtained results to support the presence of such systematic errors resulting from the binning process. Our results show that the straightforward histogram method underestimates the values of entropy with an error that is linearly proportional to the ratio of number of used bins to the size of the dataset. Also, from experiments using artificial datasets with given correlation, we find that the AMI is overestimated particularly at small dataset sizes with low and medium correlation.

From experiments done using the present introduction of probability distribution smoothing, we compare between different smoothing functions and conclude that median filter smoothing gives best efficiency in terms of the amount of

error reduction in entropy and AMI as well as complexity of computation. Results of our experiments show that probability density smoothing significantly reduces the AMI errors. Moreover, the results show that the median filter smoothing significantly reduces the errors in AMI over the greater part of the range of correlations. Preferred features in this method are its lower complexity and that it does not remove genuine correlations over most of the correlation values.

As an application of the present methodology for computing AMI, we considered the problem of reconstructing real multivariate time series from their dominant Independent Components (IC's). For this purpose, we have selected 4 real financial series representing the daily exchange rates of USD versus other currencies. Present experiments to reconstruct these financial series from their computer IC's show that significantly better reconstruction is obtained by using our algorithm with probability distribution smoothing.

REFERENCES

[1] P. D'haeseleer, S. Liang, and R. Somogyi, "Genetic network inference: from co-expression clustering to reverse engineering", Bioinformatics, vol. 16, pp.707–726, 2000.

[2] C.E. Shannon, "A mathematical theory of communication", The Bell System Technical Journal, vol. 27, pp.379–423. ibid, pp.623–656, 1948.

[3] R. Steuer, J. Kurths, C.O. Daub, J. Weise and J. Selbig, "The mutual information: Detecting and evaluating dependencies between variables, Bioinformatics, vol. 18, Suppl. 2, pp.S231–S240, 2002.

[4] D. Gencaga, N. K. Malakar, and D. J. Lary, "Survey On The Estimation Of Mutual Information Methods as a Measure of Dependency Versus Correlation Analysis", AIP Conference Proceedings 1636, 80 (2014); https://doi.org/10.1063/1.4903714

[5] Lu, "Measuring dependence via mutual information," M.S. Thesis, Queen's University, Canada, 2011.

[6] J. E. Hudson, "Signal processing using mutual information," IEEE Signal Processing, vol. 23, Issue 6, pp.50-54, 2006.

[7] K. H. Knuth, D. Gencaga, and W. B. Rossow, "Information-Theoretic methods for identifying relationships among climate variables", Earth-Sun Systems Technology Conference, 2008.

[8] A.M. Fraser, and H.L Swinney, "Independent coordinates for strange attractors from mutual information", Phys. Rev. A, vol. 33, pp.2318–2321, 1986.

[9] R. Thomas, N. Moses, E. Semple, and A. Strang, "An efficient algorithm for the computation of average mutual information: Validation and implementation in Matlab", Journal of Mathematical Psychology, vol. 61, pp.45-59, 2014

[10] A. Grinstead https://www.mathworks.com/matlabcentral/fileexchange/10040-average-mutual-information, 2006

[11] C. J. Cellucci, A. M. Albano, and P.E. Rapp, "Statistical validation of mutual information calculations Comparison of alternative numerical algorithms", Phys. Rev. E, vol. 71, pp. 066208-1-066208-14, 2005

[12] P. Qiu, A.J. Gentles, and S.K. Plevritis, "Fast calculation of pairwise mutual information for gene regulatory network reconstruction", Computer Methods and Programs in Biomedicine, vol. 94, pp.177–180, 2009

[13] A. Goneid, "Moment Preserving Approximation of Independent Components for the Reconstruction of Multivariate Time Series", International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), ISSN: 2321-8169, vol. 5, Issue 10, pp.41 - 48, 2017

[14] W. Antweiler, University of British Colombia, Sauder School of Business, http://fx.sauder.ubc.ca/data.html, 2019

[15] A. Goneid, and A. Kamel, "Reconstruction of time series using optimal ordering of ICA components", International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), ISSN: 2321-8169, vol. 5, Issue 7, pp.297 – 305, 2017