# TIME-BASED CLUSTERING AND ITS RELATIONSHIP WITH MUTUAL INFORMATION THEORY

Charles R. Tolle, *
Department of Electrical & Computer Engineering
Utah State University
Logan, Utah 84322

May 23, 1997

## Abstract

This paper re-introduces the concepts of Time-Based Clustering (TBC). Also, the ideas of over-sampling and embedding time are introduced in connection with mutual information theory. These concepts are then extended through the use of the Time-Based Clustering (TBC) problem. Mutual information curves for the Rössler system are shown to match a slice through the rich cost function space spanned by the Time-Based Clustering (TBC) solution. In closing, some possible repercussions of this find are discussed.

## 1 Introduction

This paper has been written using information obtained within an extended research project in determining the origins of generic signals[1]-[3]. The concept of determining the origin or motivation of a generic signal is very important to an engineer today. For example, when a new cooling system or aircraft is designed and tested there are times when these systems develop unexpected turbulent flows. The engineer would like to know if these flows are created by deterministic processes or if they are caused by random effects. Moreover, if the flows are deterministic based, the engineer most likely has a better chance to make an engineering change to overcome the deterministic cause of the unexpected turbulance or in short control it.

There are many methods that can be attempted for determining the origin of a signal, they include: First Return Maps, Statistical based calculation, attractor reconstruction, and a new hybrid method created by the authors[3]. Many of these methods rely on the choice of the optimal embedding time step. The optimal embedding time step has an important connection to the process of coarsening or converting a complicated signal into a symbol train used in the authorś new method. This connection is discussed further within this paper. However, the idea of embedding time step optimal sample time should be of interest for all engineers. As with many young engineers, the author was taught that under-sampling of a signal was a very big problem. Moreover, if a signal is under-sampled, information is lost within the signal. In addition, if proper low pass filtering has not been accomplished on the under-sampled signal aliasing effects come into play. However, as a young engineer the effects of over-sampling a signal were not discussed with as much importance, if at all. Over-sampling is as big a problem in attempting to learn about the dynamics of a system as

* Utah State University Rocky Mt. NASA Space Grant Consortium Fellow, TARDEC PhD. Co-op, PhD. candidate and Research Assistant for the Department of Electrical and Computer Engineering Utah State University, and Research Engineer for the Utah State University Center for Self-Organizing Intelligent Systems.

the under-sampling problem. What is meant by over-sampling? A simple example explains the concept best. Consider sampling a 1Hz sine at 1000Hz. If one now looks at the resulting digital signal not much changes from one time step to the next. In fact, as one tries to find the dynamics within the digital signal at least 1000 points are required to see the global 1Hz sine wave. In short, the general dynamics from time step to time step have been washed out by the over-sampling.

A traditional method for determining the embedding time step/optimal sample time has been based on mutual information theory. In this paper, the traditional mutual information theory solution has been shown (using the Rössler system) to match a new Time-Based Clustering (TBC) solution. Moreover, the traditional mutual information theory is a macro solution, while the TBC solution contains both the macro and micro level optimal sample time information. These ideas are discussed further within this paper.

## 2 The Rössler System

First, the reader should be come acquainted with the Rössler system. This system is a simple three state system defined as:

$$\dot{x}(t) = -(y(t) + z(t)) \tag{1}$$
$$\dot{y}(t) = x(t) + ay(t) \tag{2}$$
$$\dot{z}(t) = b + z(t)(x(t) - c) \tag{3}$$

This system under goes some important changes when the parameters $a$ and $b$ are fixed at 0.2 and the parameter $c$ is varied. As the parameter $c$ is varied from 2 to 5.7 an infinite number of period doublings occur[4] within its attractor. These period doubling events produce the strange attractor shown in Figure 1. Note that the Rössler system with the parameters $a = 0.2$, $b = 0.2$, and $c = 5.7$ exhibits chaos. Furthermore, this simple chaotic system allows for a detailed comparison between mutual information theory and the TBC problem, as will be shown shortly.

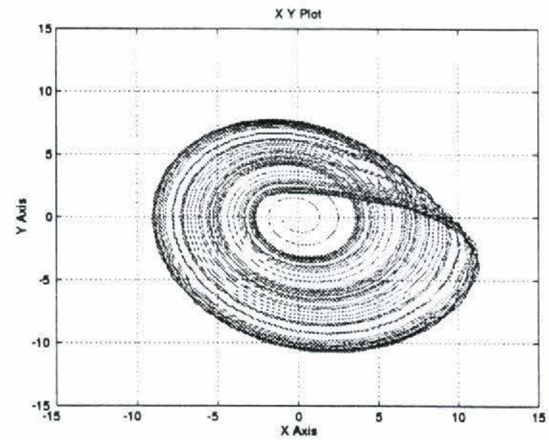


Figure 1: $x$ state versus $y$ state plot of the Rössler system time evolution ($a = 0.2$, $b = 0.2$, and $c = 5.7$).

## 3 Time-Based Clustering

This is a simple reintroduction of the concepts behind the Time-Based Clustering (TBC) problem that was originally introduced in [1]. For a more complete coverage of the TBC problem the following papers and documents are suggested [1], [2], and [3] for reading. The TBC problem has been initially designed as a coarsener used within the symbol string creation process for study of strange signals. The TBC problem is built upon the Fuzzy-C Means (FCM) clustering problem[5]-[7],[1],[3]. The formulation of the TBC problem is a generalized of the FCM with a twist. One of TBC's goals is to convert a $n$ dimensional digital signal into a one dimensional symbol train that still retains as much dynamical information contained within it as possible. One might first consider implementing time embedding practices within the data to be clustered in order to accomplish the goal. Instead, one might consider changing the idea of how a cluster is defined. Using the later idea and knowing that we are interested in time series data, it is logical that our data clusters should also contain some kind of time series information. However, one does not in general know much about the time series data that is being studied. So, in some way TBC should be allowed to determine the best way to involve the time series information within the clusters. With this notion in mind,

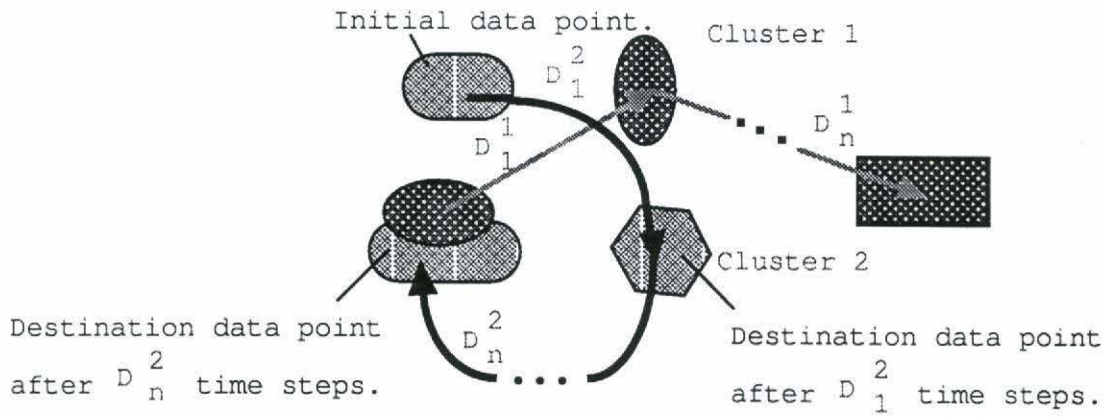


Figure 3: The generalized new cluster concept used within the new TBC problem and coarsener.

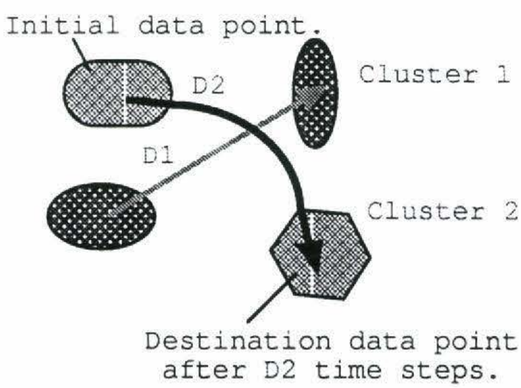


Figure 2: The new cluster concept used within the new TBC problem and coarsener.

consider a new type of cluster shown in Figure 2. Note, that by using this new concept of a cluster, a data point is only considered to be part of a cluster if it both starts in the vicinity of a cluster center and a later data point ends within the vicinity of the ending cluster center. It is easy to see that this is just a simple extension of older clustering concepts. Moreover, this simple single time step cluster idea can be generalized further to include multiple future cluster centers, see Figure 3. This extended TBC problem will be discussed further in the conclusions of this paper, as well as in [3].

By extending the cluster concept to include future dynamic data, the clusters are able to retain more dynamical information within the symbol trains that are created from TBC. That is the goal! In fact, as will be discussed later, the TBC problem solves the over-sampling problem in a manor similar to how mutual information theory solves over-sampling. So, how might one state the new TBC problem. Consider:

$$\min_{\substack{v1_i, v2_i \\ u_{ik}, D_i}} \sum_{i=1}^{C} \sum_{k=1}^{N-\Delta_{max}} u_{ik}^2 \left\| \begin{array}{ccc} x_k & - & v1_i \\ x_{k+D_i} & - & v2_i \end{array} \right\|^2 \quad (4)$$

$$W.C. \quad \begin{array}{c} u_{ik} \in [0,1] \\ \sum_{i=1}^{C} u_{ik} = 1 \ \ \forall k = 1, 2, \ldots, N \\ D_i \in [1, 2, \ldots, \Delta_{max}] \ ; \ \Delta_{max} \ll N \end{array} \quad (5)$$

where, $C$ is the number of clusters, $N$ is the number of data vectors $\{ x_k \}$ being clustered, $u_{ik}$ is the membership value of the $k^{th}$ data vector in the $i^{th}$ cluster, and $[v1_i^T v2_i^T]^T$ is center (mean) vector of the $i^{th}$ cluster. Note that the center vector holds both locational as well as dynamic signal direction information. For this paper the $\| \bullet \|$ is the Euclidean norm, however an adaptive matrix norm could be used in the future[6]-[7]. It is important to note that this problem is very difficult! Why? The fact that the $D_i$ terms are allowed to change implies that you cannot simply apply a Lagrange multiplier solution to this problem. In short you can't create a simple one step solution process such as the FCM problem has[1],[3]! Solution methods are discussed in [3]. However, note that if the $D_i$'s are held fixed, then the TBC problem reverts to the more simple FCM problem. This is a key to its solution as well as a newer definition of the TBC problem that follows. Note that the above definition seems reasonable. However, as was discovered during the research process, a better TBC problem is stated as:

$$\operatorname*{\substack{first\\local\\max}} \min_{\substack{v1_i, v2_i\\u_{ik}, D_i}} \sum_{i=1}^{C} \sum_{k=1}^{N-\Delta_{max}} u_{ik}^2 \left\| \begin{matrix} x_k - v1_i \\ x_{k+D_i} - v2_i \end{matrix} \right\|^2 \tag{6}$$

$$W.C. \quad \begin{matrix} u_{ik} \in [0,1] \\ \sum_{i=1}^{C} u_{ik} = 1 \quad \forall k = 1, 2, \ldots, N \\ D_i \in [1, 2, \ldots, \Delta_{max}] \; ; \; \Delta_{max} \ll N \end{matrix} \tag{7}$$

This might seem strange to some readers, however by maximizing over the minimized FCM TBC problems, one solves the problem of over-sampling as well as clusters the data. Over-sampling is a problem few engineers consider, however it is important in this work. The idea of choosing the perfect sampling time is very important. One knows that if the time is not fast enough information is lost! Likewise, if one samples too fast, dynamical information from one time step to another is lost. Mutual information theory attempts to lessen this problem. Basically, one attempts to find a time step within a sampled signal such that the information within the initial time step does not tell much about the future time step. However, we want that time step to be as short as possible so that we do not under-sample the signal. With this in mind, from studies in mutual information theory, it has been assumed that the best choice for the subsample time step is where the first local minimum occurs in the mutual information curve[8]. If such a minimum does not exist, then the choice of no subsampling should be made[8]. The mathematical concept behind this is that one is attempting to decorrelate the information from one time step to the next. Explained slightly different, one is attempting to spread the data points out but not so far that one losses information, just far enough to get rid of redundant information. It will been shown that this is what the above restated TBC problem accomplishes! Furthermore, note that the first TBC problem statement should be used when there is not a local minimum on the mutual information curve – or likewise a local maximum on the TBC constant $D_i$ curve. The lack of a local minimum occurs in many systems, in fact it always occurs within simple map based chaotic systems. These concepts are discussed more thoroughly later in [3]. Moreover, at this point it should be noted that the second TBC problem solves the mutual information problem over a larger solution space. This is due to the the ability of the clusters to localize the mutual information across the data space. The TBC problem thus finds the best subsample time steps for each cluster separately. In short mutual information theory attacks the global problem while TBC can solve both the global as well as the local over-sampling problem. One can think of the TBC problem as a generalized mutual information problem. Extending this idea further, as is shown latter, mutual information theory produces a simple curve that is searched over for the best solution, while TBC method produces a much higher dimensional surface that must be searched for the best result. In short, TBC problem could be thought of as a multiple localized mutual information solution. This is shown in the next section.

# 4 Mutual Information and Time-Based Clustering

How is TBC and mutual information theory connected? In order to answer this question, one must first know what mutual information theory is useful for and how it is implemented. Mutual information theory is used in reconstructing attractors. Mutual information is a method for deciding on the optimal time embedding step used within the reconstruction process. As was discussed earlier, this choice of the embedding time step is closely tied to the idea of optimal sampling time. In other words, traditional sampling theory gives one the maximum sample time needed so that the signal can be sampled and then reconstructed. Likewise, mutual information theory addresses the over-sampling problem. It is an attempt to set the minimum sample time step size. Moreover, as the maximum sample time addresses aliasing, the minimum sample time addresses the loss of dynamical information within the sampled signal. Therefore, by connecting mutual information theory with the TBC problem, a claim can be made: by using TBC as the means to produce a symbol train, one gains over-sampling protection as well as clusters that contain more of the original signals dynamical properties, as discussed earlier. The tie to dynamical properties comes from the idea that a point is only a member of a cluster if it starts within a traditional cluster and evolves in time to a traditional final cluster.

The mutual information process for choosing the optimal sample time is simple to understand. The concept is based on choosing a sample time that has as short a time lag as possible but also produces the minimum amount of information content about the next time sample from the current time sample. Moreover, the optimal sample time should maximize the dynamical information in between each sampled time step. This is done by choosing the minimum informational content time step on the informational curve. The reason for choosing the minimum informational content is that we want the data to change as much as possible in between each time sample but still be relatively close in time. This concept is the oversampling idea. However as was discussed above, there is a special case. If the mutual information curve does not contain any local minimum then the choice of the time step sample should be 1[8]. This occurs for all simple mapping functions like the tent map, Lozi system, etc [3].

The mutual information curve can be calculated with the following equation:

$$I(T) = \sum_{b}^{M} \sum_{b'}^{M} P_{bb'}(T) \log_2 \left[ \frac{P_{bb'}(T)}{P_b(0) P_{b'}(T)} \right] \quad (8)$$

where the probabilities $P_b(0)$, $P_{b'}(T)$ and the joint probabilities $P_{bb'}(T)$ are calculated from the uniformly sampled/coarsened signal. Moreover, $P_i(K)$ is defined as the probability of symbol $i$ within the sampled signal, $s(n+K)$. Also, $P_{ij}(K)$ is defined as he probability of symbol $i$ being followed $K$ steps later by symbol $j$ within the symbol train $s(n)$. Note that this notation is not standard, however it is hoped that it will limit confusion within the above equation. Also note, that there is a minor typographical error within Abarbanel's paper[8] on how to calculate the mutual information curve. Continuing on, the mutual information curve is calculated by first uniformly sampling/coarsening the signal into $M$ bins, see Figure 4. Once the signal has been made into a simple histogram, see Figure 5, the probability of each symbol $(P_i(K))$ can be easily calculated by dividing the number points within each bin by the number of points within the signal. Notice that as the symbol string grows in length the $P_i(K)$ becomes constant for all choices of $K$. Therefore, if the symbol string is long then we need only calculate the histogram in Figure 5 once. Thus, we do not need to calculate the $P_i(K)$ for each time sifted (by $K$) signal. Likewise, $P_{bb'}(T)$ is calculated by creating a 2 dimensional histogram for symbol pairs separated by $T$ time steps within the uniform coarsened symbol train, see Figure 6. Once the probabilities have been calculated the mutual information curve is simple to create. This has only been an overview of mutual information, if the reader wants a more detailed coverage I suggest reading Abarbanel's paper [8].

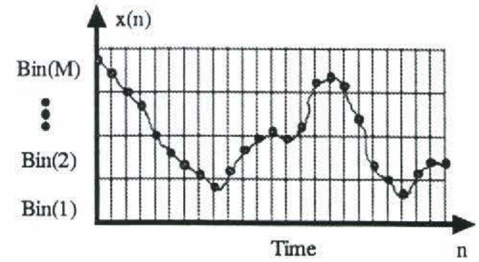


Figure 4: Example of uniform quantinization used within the calculation of mutual information.

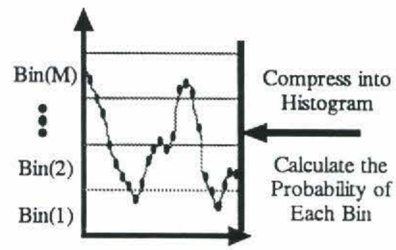


Figure 5: A method for calculating symbol probability used in the calculation of mutual information.

Using the methods above, the mutual information curve for the well known Lorenz system[4],[3] was calculated. It was found that the Lorenz's mutual information curve had an inverse relationship to the cost function obtained from the TBC problem, see Figure 7. This re-

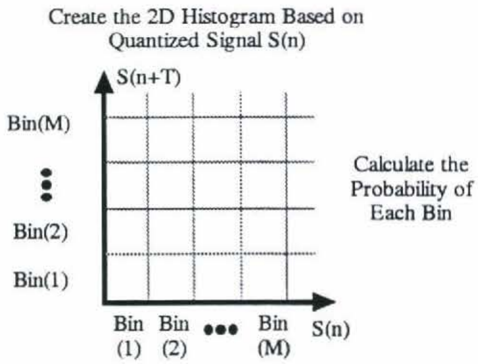


Figure 6: Example of the joint symbol probability calculation used within the mutual information calculation.

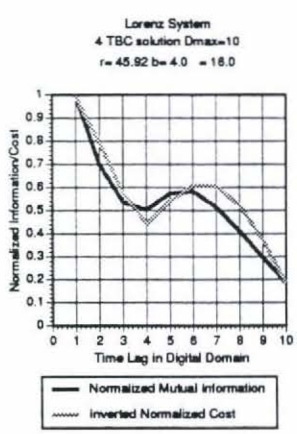


Figure 7: A plot of the normalized Lorenz mutual information curve and the normalized inverse constant $D$ cost function from the TBC problem.

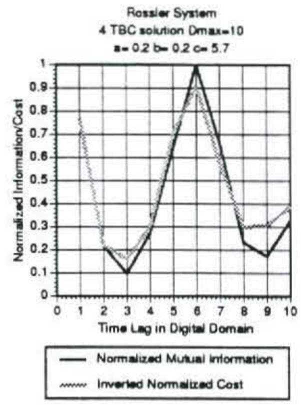


Figure 8: A plot of the normalized Rössler mutual information curve and the normalized inverse constant $D$ cost function from the TBC problem (4 bins).

lationship occurs for a very special case of the TBC problems. This special case is where the values of the $D$'s are held constant for each cluster within the clustering problem. This in fact is the mutual information's time step choice analog within the TBC problem. By holding the $D$'s constant for each cluster, we are simulating the use of a global embedding time step within the problem. Notice that the embedding time step is what mutual information theory is designed to find. Thus by comparing these two curves once normalized they should match. The inverse relationship comes from the idea that the clustering problem penalizes for informational mismatches with a higher cost, while mutual information theory gives a lower value for information mismatches. Therefore, we should expect the inverse relationship and that is what has been found.

Wait! This isn't the whole story. By using the Rössler system as an example, the concepts of mutual information theory can be extended into a local solution instead of the global solution it gives currently. What does this mean? Mutual information theory allows one to only calculate a global information relation between time steps. Likewise, by fixing the $D$'s to be constant only the global mutual information calculation is obtained by using TBC problem. This is shown in Figure 8. However, if the $D$'s are allowed to vary, local informational calculations can be obtained

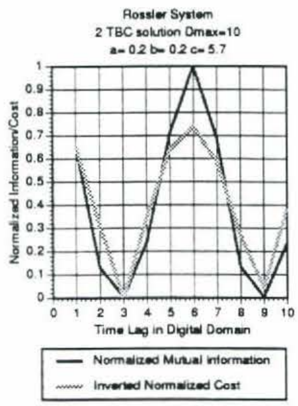


Figure 9: A plot of the normalized Rössler mutual information curve and the normalized inverse constant $D$ cost function from the TBC problem (2 bins).

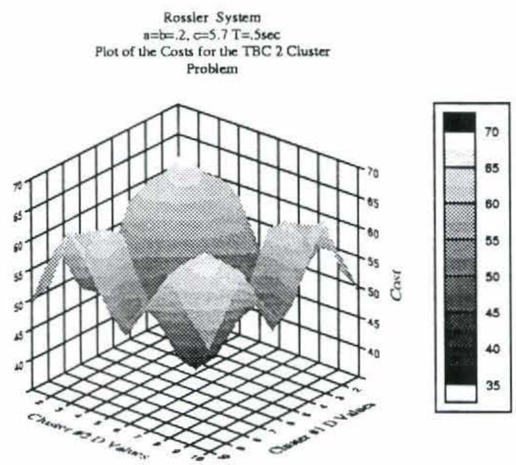


Figure 10: A plot of the costs for the Rössler TBC problem for 2 clusters.

using the TBC problem. It should be noted in the Figures 7 and 8, the number of bins was held to 4 for the mutual information calculation and 4 clusters where used in the TBC problem. This is an attempt to make the comparisons as valid as possible. In this light, the Rössler system is simple enough, that useful mutual information curves for 2 bins is possible, see Figure 9. This allows us to use only 2 clusters in the TBC problem. By using only 2 clusters a plot of the whole TBC solution surface can be shown, see Figure 10. As before this cost surface can be normalized and inverted to compare it with the expected results from mutual information theory, see Figures 11 and 12. If we compress the map into 2

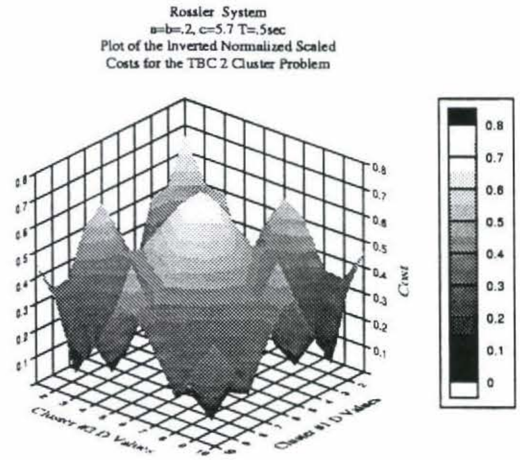


Figure 11: View one of the normalized inverted costs for the Rössler TBC problem for 2 clusters.

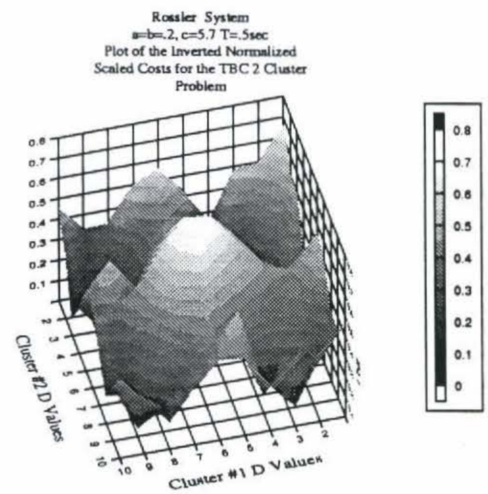


Figure 12: View two of the normalized inverted costs for the Rössler TBC problem for 2 clusters.

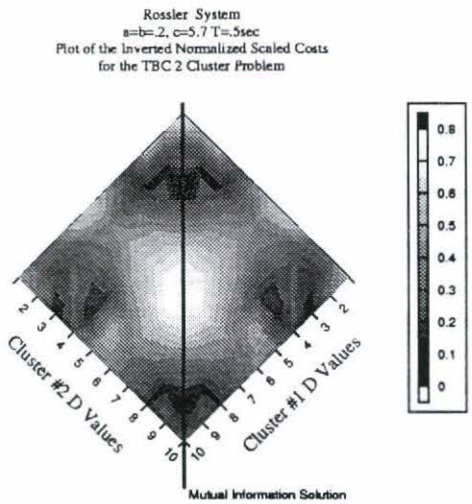


Figure 13: Top view of the normalized inverted costs for the Rössler TBC problem for 2 clusters.

dimensions, see Figure 13, the standard mutual information curve solution can be shown within this more rich localized TBC solution, i.e. the traditional mutual information solution is along the marked diagonal in Figure 13. The ability of the TBC problem to extended the ideas of mutual information adds greatly to the advantages of using it for symbol string creation.

## 5 Conclusions and Future Directions

The main purpose of this paper was to show that Time-Based Clustering(TBC) and mutual information theory are connected. This connection was shown through experimental comparisons between the TBC cost function results and the mutual information curves calculation for the Lorenz and Rössler systems. Moreover, the experimental relationship was discussed in terms of the underlining mathematical parallels within each problem. The importance of both TBC and mutual information theory to obtain the optimal sample time was discussed briefly. However, the use of optimal sample time is a separate discussion outside of what is proper in this paper. More importantly, by showing the connection between TBC and mutual information theory, a more generalized concept of a mutual information surface is obtained. Furthermore, by considering the more advanced generalized TBC problem, see Figure 3, new more interesting optimal attractor reconstruction could be attempted. These new reconstructions could be based on different embedding time steps used within the data embedding process. This concept is not readily available within the mutual information theory at the present. A more detailed discussion of these concept can be read in [3]. The ideas presented in this paper are just a jumping off point. Additional research is needed in this area to exploit this new generalization of the mutual information concepts.

## References


[1] Charles R. Tolle, "Chaos and Control", Proceedings of the 1995 Rocky Mountain NASA Space Grant Consortium Conference, June 1995.

[2] Charles R. Tolle, "A PhD Proposal on Chaos and Control", Utah State University PhD. Proposal, Sept. 1995.

[3] Charles R. Tolle, "Time-Based Clustering and Its Application to Determining a Signal's Origin: Deterministic Chaos or Random", Utah State University PhD. Dissertation, currently being edited, 1997.

[4] Steven H. Strogatz, Nonlinear Dynamics and Chaos - With application to Physics, Biology, Chemistry, and Engineering, Addison-Wesley Publishing Company, Reading, Main, 1994.

[5] Robert L. Cannon, Jitendra V. Dave, and James C. Bezdek, "Efficient Implementation of Fuzzy c-Means Clustering Algorithms," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-8, No. 2, 1986, pp. 248-255.

[6] R. W. Gunderson, "Adaptive FCV clustering algorithm," Int. J. Man-Machine Studies, Vol. 19, 1983, pp. 97-104.

[7] Donald E. Gustafson, and William C. Kessel, "Fuzzy Clustering With a Fuzzy Covariance Matrix," Proc. IEEE CDC, San Diego, CA, Jan. 10-12, 1979, pp. 761-766.

[8] Henry D.I. Abardanel, Reggie Brown, John J. Sidorowich, and Lev Sh. Tsimring, "The analysis of observed chaotic data in physical systems," Reviews of Modern Physics, Vol. 65, No. 4, October, 1993, pp. 1331-1392.