SOM-BASED SIMILARITY INDEX MEASURE: QUANTIFYING INTERACTIONS BETWEEN MULTIVARIATE STRUCTURES

Anant Hegde, Deniz Erdogmus, Yadunandana N. Rao,
Jose C. Principe, Jianbo Gao

Computational NeuroEngineering Laboratory,
Electrical & Computer Engineering Department
University of Florida, Gainesville, FL 32611.
Email: [ahegde,deniz, yadu, principe] @cnel.ufl.edu, gao@ece.ufl.edu

Abstract. This paper addresses the issue of quantifying asymmetric functional relationships between signals. We specifically consider a previously proposed similarity index that is conceptually powerful, yet computationally very expensive. The complexity increases with the square of the number of samples in the signals. In order to counter this difficulty, a self-organizing map that is trained to model the statistical distribution of the signals of interest is introduced in the similarity index evaluation procedure. The SOM based technique is equally accurate, but computationally less expensive compared to the conventional measure. These results are demonstrated by comparing the original and SOM-based similarity index approaches on synthetic chaotic signal and real EEG signal mixtures.

INTRODUCTION

Understanding the interrelations between multiple time-series has numerous applications in signal processing and engineering. A key aspect of understanding how a system works is, understanding how information at its different nodes is coupled and how it propagates through these nodes. In particular, quantifying the interactions between the signals at various channels of an EEG recording across the temporal lobe could potentially help us predict epileptic seizures. One can also trace the epileptic foci by understanding various coupling characteristics between EEG traces at different time-instances. Various linear and nonlinear techniques have been developed to quantify the degree of synchronization. Cross-correlation analysis is one of the earliest and most relied-on linear techniques in this effort. Directed coherence, direct transfer functions (DTF) and partial directed coherence (PDC) [1-3] are amongst the other methods that have been lately proposed and researched on. However, these approaches describe only the linear structural inference between the stochastic processes. Linearity assumptions restrict the applicability, because most of the real world signals, including EEG, are generated from non-linear interactions between complex systems.

Nonlinear dependencies between multiple signals have also been studied in the last two decades, but with limited success. Popular methods utilize concepts based

on generalized mutual information [4], and instantaneous phase measures using Hilbert transforms [5,6] and Wavelet transforms [7]. The difficulty with these methods has been the need to use very long data series and computational complexity due to the handling of this data. Additional requirements, such as narrow-band signals, also hindered the general applicability of some methods. Eckmann *et al.* [8] proposed the method of recurrence plots (RPs) that represents the recurrence of states in the phase-space trajectory of a chaotic signal. Since chaotic systems are non-linear in nature, this method has been fairly successful in detecting bifurcations and non-stationarities in time sequences [9]. Cross-recurrence plot (CRP) is an extension of the RP idea to multi-dimensional time signals [10]. The CRP has found use in describing the time-dependency between multiple time-series recorded from multiple locations. However, the lack of quantitative information and the computational complexity makes it tedious for analyzing large sets of data. One of the common drawbacks of most of the measures is that they fail to indicate the direction of information flow (or influence). In a closed-system, it is reasonable to expect that there exist linear or nonlinear dependencies between the signals acquired from measurements at various points. In some engineering applications, such as the prediction of epileptic seizures, it is essential to identify the information flow direction between these multiple nodes in the system.

Recently, Arnhold *et al.* [11] introduced the similarity–index technique (SI) to measure such asymmetric dependencies between time-sequences. Conceptually, this method relies on the assumption that if there is a dependency between two signals, the neighboring points in time will also be neighboring points in state space. Since this requires searching for the nearest neighbors in the state space (formed by embedding) for large data sets, the computational complexity becomes unmanageable. In this paper, we propose a self-organizing map (SOM) based improvement to this method to reduce computational complexity, while maintaining accuracy. This is achieved by mapping the embedded data from signals onto a quantized output space through a SOM specialized on these signals, and utilizing the activation of SOM neurons to infer about the influence directions between the signals, in a manner similar to the original SI technique.

THE SIMILARITY INDEX TECHNIQUE

In this section, brief description of the original SI measure is provided. Given two signals, X and Y, the similarity index, which is defined as

$$S(X \mid Y) = \frac{1}{N} \sum_{n=0}^{N-1} \frac{R^n(X)}{R^n(X \mid Y)} \tag{1}$$

quantifies the average influence of Y on X. Here, $R^n(X)$ measures the average Euclidean distance between the sample-vector x_n, which is constructed by embedding the original time series in a delay vector, and its k nearest neighbors in a neighborhood of radius ε, at time instant n. Similarly, the quantity $R^n(X \mid Y)$

measures the average Euclidean distance between x_n and the sample-vectors of X whose time indices correspond to the time indices of the nearest neighbors of y_n.

By definition, $0 \le R^n(X) \le R^n(X \mid Y)$, and the ratio in (1) is always in [0,1]. As a consequence, $S(X \mid Y) = 1$ implies X is completely dependent on Y. This suggests that recurrence of a state in Y implies a recurrence in X [12]. On the same principles, $S(X \mid Y) = 0$ implies complete independence between X and Y. Similarly, it is possible to quantify the average dependence of Y on X by

$$S(Y \mid X) = \frac{1}{N} \sum_{n=0}^{N-1} \frac{R^n(Y)}{R^n(Y \mid X)} \tag{2}$$

Comparing $S(X \mid Y)$ and $S(Y \mid X)$, we can determine which signal is more dependent on the other. By design, the similarity index can identify nonlinear dependencies.

The difficulty with this approach is that at every time instant, we must search for the k nearest neighbors of the current embedded signal vectors among all N sample vectors; this process requires $O(N^2)$ operations. In addition, the measure depends heavily on the free parameters, namely, the number of nearest neighbors and the neighborhood size ε. The neighborhood size ε needs to be adjusted every time the dynamic range of the windowed data changes.

SOM-BASED SIMILARITY INDEX

The SOM-based SI algorithm is designed to reduce the computational complexity of the SI technique. The central idea is to create a statistically quantized representation of the dynamical system using a SOM [13,14]. For best generalization, the map needs to be trained to represent *all* possible states of the system (or at least with as much variation as possible). As an example, if we were to measure the dependencies between EEG signals recorded from different regions of the brain, it is necessary to create a SOM that represents the dynamics of signals collected from all channels. The SOM can then be used as a prototype to represent any signal recorded from any spatial location on the brain, assuming that the neurons of the SOM have specialized in the dynamics from different regions.

One of the salient features of the SOM is topology preservation [13,14]; i.e., the neighboring neurons in the feature space correspond to neighboring states in the input data. In the application of SOM modeling to the similarity index concept, the topology preserving quality of the SOM enables us to identify neighboring states of the signals by neighboring neurons in the SOM.

Assume that X and Y are two time series generated by a system, which are embedded into two vector signals in time using delays. Define the activation region of a neuron in the SOM as the set of all input vectors (the embedded signal vectors) for which the neuron is the winner based on some distance metric (Euclidean in most cases). Let X_n be the set of time indices of input vectors x_j that are in the activation region of the winner neuron corresponding to the input vector

x_n at time n. Similarly define the set Y_n. Then the procedure to estimate the directed similarity indices between X and Y using a SOM is as follows:

1. Train a SOM using embedded vectors from both X and Y as the input.
2. At time n, find W_n^x, the winner neuron for vector x_n, and find W_n^y, the winner neuron for vector y_n.
3. Determine the sets X_n and Y_n for W_n^x and W_n^y, respectively.
4. Determine the nearest neurons $W_{n,j}^y$ corresponding to vectors y_j, where $j \in X_n$. Determine the nearest neurons $W_{n,j}^x$ corresponding to vectors y_j, where $j \in Y_n$.
5. Calculate $R^n(X \mid Y) = (1/q)\sum_{r=1}^{q} \left\| W_n^x - W_{n,j}^x \right\|$, where q is the number of elements of X_n. Calculate $R^n(Y \mid X) = (1/q)\sum_{r=1}^{q} \left\| W_n^y - W_{n,j}^y \right\|$, where q is the number of elements of Y_n.
6. Find $R(X \mid Y)$ as the average of $R^n(X \mid Y)$ over all n. Find $R(Y \mid X)$ as the average of $R^n(Y \mid X)$ over all n.
7. Compute the normalized similarity index as

$$\chi = \frac{R(Y \mid X) - R(X \mid Y)}{R(X \mid Y) + R(Y \mid X)} \tag{3}$$

By construction, large values of $R(X \mid Y)$ and $R(Y \mid X)$ imply weaker dependency or no dependency. The normalized similarity index χ, on the other hand can point out directed influences between the two signals. Specifically, positive values of χ indicate an influence of Y on X, while negative values indicate the opposite.

Higher the magnitude of χ indicates a stronger coupling of the signals in the direction indicated by the sign. When χ is close to zero, an ambiguity occurs, since the two signals could be independent or coupled to each other equally in both directions. This ambiguity can be resolved by observing the individual values of $R(X \mid Y)$ and $R(Y \mid X)$.

The computational savings of the SOM approach is an immediate consequence of the quantization of the input (signal) vector space. The search for nearest neighbors will involve $O(Nm)$ operations as opposed to the $O(N^2)$ of the original algorithm, where N is the number of samples and m is the number of neurons in the SOM ($m<<N$ by design).

SIMULATION RESULTS

In this section, we demonstrate the viability of the SOM-based similarity index approach in determining couplings and influence directions between synthetic and

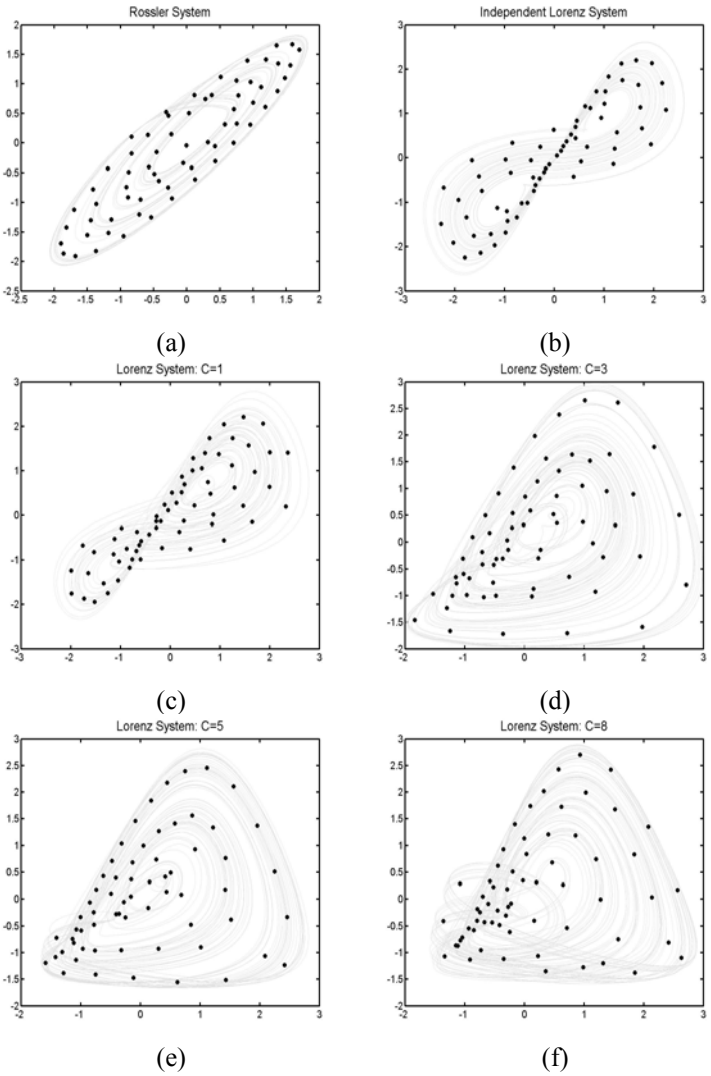

Figure 1. Phase-space trajectories of the Rossler-Lorenz system for various coupling strengths a) Rossler b) Lorenz (C=0) c) Lorenz (C=1) d) Lorenz (C=3) e) Lorenz (C=5) f) Lorenz (C=8). The SOM weights (dots) for each signal are superimposed on the trajectory.

real signals. One case study considers a coupled Rossler-Lorenz system (as described in [12]), and the other considers real EEG signals.

Rossler-Lorenz signals: The same Rossler-Lorenz example used by Quiroga *et al.* [11] is used here. A synthetic nonlinear dependency between a Rossler (X) and

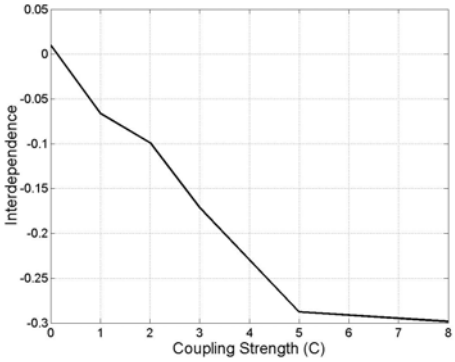

Figure 2. The normalized similarity index versus coupling strength for the Rossler-Lorenz system.

a Lorenz (Y) system is created by having the second state of the Rossler system drive the Lorenz system in the following manner:

$$\begin{array}{ll} \text{Roessler} & \text{Lorenz} \\ \dot{x}_1 = -6\{x_2 + x_3\} & \dot{y}_1 = 10(-y_1 + y_2) \\ \dot{x}_2 = 6\{x_1 + 0.2x_2\} & \dot{y}_2 = 28y_1 - y_2 - y_1 y_3 + Cx_2^2 \\ \dot{x}_3 = 6\{0.2 + x_3(x1 - 5.7)\} & \dot{y}_3 = 10(-y_1 + y_2) \end{array} \quad (4)$$

where C is the coupling strength. Two SOMs, one corresponding to the Rossler system and the other for the Lorenz system (a new one for each value of C), were trained separately on embedded data (using an embedding delay of 0.3 time-units and embedding dimension of 4) from these two signals. The phase-space dynamics of the Rossler system and the Lorenz system (for different C values) are shown in Fig. 1 along with the SOMs that are trained on their corresponding embedded vectors.

Each SOM is an 8x8 rectangular grid, and is trained on a set of 4000 samples using a Gaussian neighborhood function for 1000 iterations. The neighborhood radius (standard deviation of the Gaussian neighborhood function) is exponentially annealed starting from an initial value of 6 with a time constant of 100. The step size is also annealed exponentially from 0.08 using the same time constant.

Using the SOM-based similarity index approach, the normalized indices χ are calculated for coupling strengths of C = 0, 1, 2, 3, 5, 8. The results are presented in Fig. 2. Negative values of χ suggest that the method deduced the influence of the Rossler system on the Lorenz system. In addition, the increasing strength of the coupling from the Rossler dynamics to the Lorenz dynamics is captured by the increasing magnitude of the normalized similarity index. The index also captures the independence of the two dynamics when C=0 by yielding very high $R(X \mid Y)$ and $R(Y \mid X)$ values (indicating independence); these values are 1.957 and 1.996, respectively. These results are in agreement with those presented by Quiroga *et al.* [12]. An interesting observation is that as C increases beyond 5, since the attractor

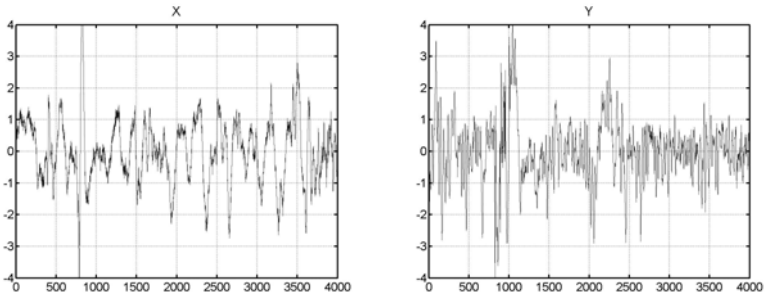

Figure 3. The original EEG signals *X* and *Y*.

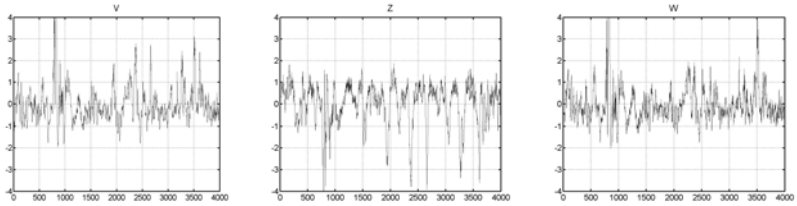

Figure 4. The nonlinearly mixed (synthetic) EEG signals *V*, *Z*, and *W*.

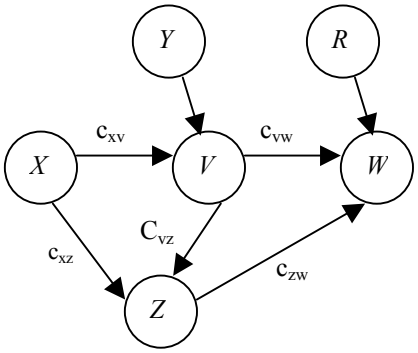

Figure 5. Schematic diagram of the coupling function between the signals.

that the driven Lorenz system follows does not change significantly, the normalized similarity index also experiences negligible change.

EEG signals: In this example, real EEG signals recorded from epileptic patients are considered. Clinically, it is useful to know or find out the direction of information flow through EEG signals. This analysis may help locate the epileptic foci of the seizures as well as providing means of predicting them.

In this signal, since the EEG measurements are generated by a closed system (the brain), we assume that the dynamical statistics of the signals observed at different channels can be modeled using a single SOM, unlike the previous

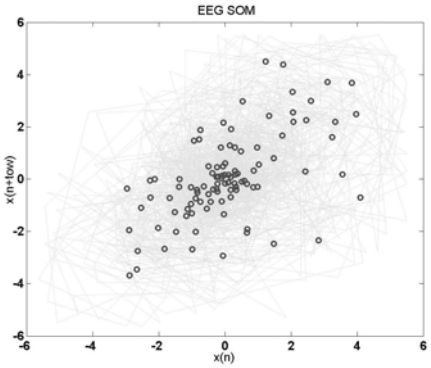

Figure 5. The phase-space trajectory of the training EEG signal (dimensions X and X) and the weights of the trained EEG-SOM (circles).

synthetic example, where a separate SOM was used for each signal. The SOM, however, must be trained using data that represents all possible psycho-physiological states that the EEG signals might exhibit. In the case of an epilepsy patient, these include pre-ictal, itcal and post-ictal states, in addition to the inter-ictal state.

In this example, EEG signals collected from two different patients at different locations (labeled X and Y) are used. A synthetic nonlinear functional relationship with influence direction from X to V, W, and Z is created according to (5). Care is taken in choosing these functions to make sure that the synthetic EEG mixtures in (5) exhibit the characteristics of real EEG signals. The signals are shown in fig. 4. This is achieved by verifying that the time structure and the power spectra of these signals are consistent with that of an EEG signal.

$$\begin{aligned} v(n) &= y(n) + c_{xv}x(n-2)x(n-3) \\ z(n) &= c_{vz}v(n-2) + c_{xz}x(n-2) \\ w(n) &= 0.05r(n) + c_{zw}z(n-2) + c_{vw}v(n) \end{aligned} \tag{5}$$

Here, $x(n)$ and $y(n)$ denote the original time sequences and $v(n)$, $w(n)$, and $z(n)$ denote the synthetic signals driven by the two original signals. In addition, $r(n)$ is a zero-mean unit-variance Gaussian noise term. The synthetic EEG signals are shown in Fig. 4 and the flow diagram representing the relationships in (5) is depicted in Fig. 5.

A 10x10 rectangular SOM (referred to as EEG-SOM) is trained using 3000 samples of embedded EEG data (with an embedding dimension of 10 and embedding delay of 30ms). The phase-space trajectory of the training data and the weights of the trained SOM are shown in Fig. 6. The normal EEG state is represented by the smaller amplitude activity (the dominant portion of the training data), whereas the larger amplitudes correspond to the spiky, sharp, and slow wave activity formed during the ictal state of the brain, or to artifacts formed due to muscle movements, etc. After training the SOM, the normalized similarity index between the original signals X and Y is evaluated to verify that these EEG signals are indeed independent.

The dependencies between X, V, W, and Z are also evaluated using the SOM to calculate the normalized similarity index. The results are summarized in Table 1, where both the coupling strength and the estimated similarity index between pairs of signals are presented.

Coupling Strength	χ		$S(X,Y) = S(X \mid Y) - S(Y \mid X)$	
$c_{xv} = 1$	-0.1668	$X->V$	-0.1112	$X->V$
$c_{xz} = 2$	-0.0756	$X->Z$	-0.0612	$X->Z$
$c_{vz} = -0.8$	0.0901	$Z->V$	0.0572	$Z->V$
$c_{vw} = 1$	-0.213	$V->W$	-0.0336	$V->W$
$c_{zw} = 0.3$	-0.1225	$Z->W$	-0.0644	$Z->W$

Table 1. Coupling strength between pairs of signals, the normalized similarity index and the original Similarity index between them.

The results obtained from the SOM-based SI measure and the original SI measure (in Table 1) is in perfect agreement. We conclude that X influences V and Z, V influences Z and W, and W influences Z. Comparing these with the flow diagram in Fig. 5, it is seen that all directional couplings are consistent with the *true* construction except for the relationship between V and Z. Possibly, this discrepancy is due to some cancellations between the couplings from X and from V. Also, we can see that V is exclusively constructed from the X and the Y components and does not have any independent oscillations of its own, unlike W.

These results indicate that the similarity index approach might not produce results that are consistent with what one would expect from the equations (if these are known) when the coupling diagram has closed loops.

CONCLUSIONS

The similarity index measure determines directional dependencies between two signals using the basic assumption that two related signals will have similar recurrences of the embedded state vector. This method has high computational complexity in terms of the number of samples, since a search for nearest neighbors must be performed in the phase-space of the signal.

In this paper, we proposed a SOM-based approach to estimate the similarity index. This approach reduces the computational complexity drastically by exploiting the accurate quantization properties of the SOM in representing the dynamics of the signal in the phase space. Another advantage of the SOM-based approach is that the difficulties that the original similarity index approach encounters in handling nonstationary data (such as the necessity to tweak parameters) are eliminated by training the SOM using samples from various regimes of the nonstationary system.

On the other hand, the SOM-based approach might suffer from inaccuracy if the quantization is severe. Therefore, the size of the SOM could be decided by a trade-off between representation accuracy and computational complexity. Future studies will address the issue of scenario-dependent SOM size selection, as well as

determining a suitable statistical normalization for the SOM-based method that will result in inferences with confidence measures. In addition, the algorithm must be modified such that it takes the possibility of having closed loops to produce results that are more consistent with what one would expect from the dynamical equations of the system.

Acknowledgment: This work is supported by NIH grant RO1 NS 39687.

REFERENCES

[1] L.A. Baccala, K. Sameshima, "Overcoming the limitations of correlation analysis for many simultaneously processed neural structures," Brain Research, vol. 130, pp. 33-47, 2001.
[2] L.A. Baccala, K. Sameshima, "Partial directed coherence: a new concept in neural structure determination", Biological Cybernetics, vol. 84, pp.463-474, 2001.
[3] M. Kaminiski, M. Ding, W.A. Truccolo, S.L. Bressler, "Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance", Biological Cybernetics, vol. 85, pp.145-157, 2001.
[4] B. Pompe, "Measuring statistical dependencies in a time series", Journal of Statistical Physics, vol. 73, pp.587-610, 1993.
[5] D. Hoyer, O. Hoyer, U. Zwiener, "A new approach to uncover dynamic phase coordination and synchronization", IEEE transactions on biomedical engineering, Vol. 47, pp.68-74, 2000.
[6] M.G. Rosenblum, J. Kurths, "Analysing synchronization phenomena from bivariate data by means of the Hilbert transform," Nonlinear Analysis of Physiological Data, (H. Kantz, J. Kurths, G. Mayer-Kress, eds.), pp. 91-99, Springer, Berlin, 1998.
[7] J.P. Lachaux, E.Rodriguez, J.Martinerie, F.Varela "Measuring phase-synchrony in brain signals", Human Brain Map. Vol. 8, pp.194-208, 1999.
[8] J.P. Eckmann, K.D. Ruelle, Europhysics Letters, vol. 5, pp. 973, 1987
[9] J. Gao, H. Cai, "On the structures and quantification of recurrence plots," Physics Letters A, vol. 270, pp. 75-87, 2000.
[10] N. Marwan, J. Kurths, "Nonlinear analysis of bivariate data with cross recurrence plots," Physics Letters A, vol. 302, pp. 299-307, 2002.
[11] J. Arnhold, P. Grassberger, K. Lehnertz, C.E. Elger, "A robust method for detecting interdependencies: Application to intracranially recorded EEG," Physica D, vol. 134, pp. 419-430, 1999.
[12] R.Q. Quiroga, J. Arhnold, P. Grasberger, "Learning driver-response relationships from synchronization patterns," Physical Review E, vol. 61, pp. 5142-5148, 2000.
[13] S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd edition, Prentice Hall, 1999.
[14] J.C. Principe, N.R. Euliano, W.C. Lefebvre, Neural and Adaptive Systems: Fundamentals Through Simulations, John Wiley & Sons, 2000.