Local Linear ICA for Mutual Information Estimation in Feature Selection

Tian Lan, Deniz Erdogmus
Department of Biomedical Engineering, OGI, Oregon Health & Science University, Portland, Oregon, USA
E-mail: {lantian,deniz}@bme.ogi.edu

Abstract – Mutual information is an important tool in many applications. Specifically, in classification systems, feature selection based on MI estimation between features and class labels helps to identify the most useful features directly related to the classification performance. MI estimation is extremely difficult and imprecise in high dimensional feature spaces with an arbitrary distribution. We propose a framework using ICA and sample-spacing based entropy estimators to estimate MI. In this framework, the higher dimensional MI estimation is reduced to independent multiple one-dimensional MI estimation problems. This approach is computationally efficient, however, its precision heavily relies on the results of ICA. In our previous work, we assumed the feature space has linear structure, hence linear ICA was adopted. Nevertheless, this is a weak assumption, which might not be true in many applications. Although non-linear ICA can solve any ICA problem in theory, its complexity and the requirement of data samples restrict its application. A better trade-off between linear and non-linear ICA could be local linear ICA, which uses piecewise linear ICA to approximate the non-linear relationships among the data. In this paper, we propose that by replacing linear ICA with local linear ICA, we can get precise MI estimation without greatly increasing the computational complexity. The experiment results also substantiate this claim.

Keywords – Local Linear ICA, Mutual Information, Entropy Estimation, Feature Selection

I. INTRODUCTION

Mutual information is an important tool in many applications, such as communications, signal processing, and machine learning. Specifically, in pattern recognition, dimensionality reduction and feature selection based on mutual information maximization between features and class labels has attracted increasing attention, because this approach can find out the most relevant features, therefore (i) reduces the computational load in real-time system; (ii) can eliminate irrelevant or noisy features, hence increases the robustness of the system; (iii) is a filter approach, which is independent of the design of classifier, and is more flexible.

The MI based method for feature selection is motivated by lower and upper bounds in information theory [2,3]. The average probability of error has been shown to be related to MI between the feature vectors and the class labels. Specifically, Fano's and Hellman & Raviv's bounds demonstrate that probability of error is bounded from below and above by quantities that depend on the Shannon MI between these variables. Maximizing this MI reduces both bounds, therefore, forces the probability of error to decrease.

One of the difficulties of this MI-based feature selection method is that, estimating MI requires the knowledge of joint pdf of the data in feature space. However, since features are generally mutually dependent, feature selection in this manner is typically suboptimal in the sense of maximum joint mutual information principle. Several MI-based methods have been developed for feature selection in the past years [4-9]. Unfortunately, all of these methods failed to solve the high dimensional situation.

In practice, the mutual information must be estimated nonparametrically from the training samples. Although this is a challenging problem for multiple continuous-valued random variables, the class labels are discrete-valued in the classification context. This reduces the problem to just estimating entropies of continuous random vectors. Furthermore, if the components of the random vector are independent, the joint entropy becomes the sum of marginal entropies. Thus, the joint mutual information of a feature vector with the class labels is equal to the sum of marginal mutual information of each individual feature with the class labels, provided that the features are independent. In our previous work, we exploited this fact and proposed a framework using ICA transformation and sample-spacing estimator to estimate the mutual information between features and class labels [1]. This framework is superior because it is open to diverse algorithms, i.e. each component, including ICA transformation and entropy estimator can be replaced by any qualified algorithm/alternative.

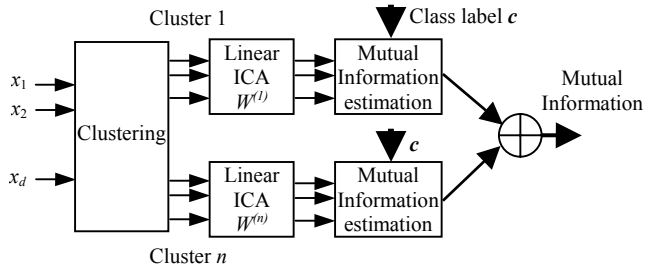

Figure 1. Local linear ICA for mutual information

We employed the cumulant-based generalized eigenvalue decomposition (GED) approach [10] to determine the linear ICA transformation and the sample-spacing estimator [11] as the marginal entropy estimator in EEG signal classification [12]. Although linear ICA yields good performance in some applications, it is a weak assumption in many real world problems. Since this framework heavily relies on the performance of ICA, the precision of MI will be greatly impaired in nonlinear situations. Recently, nonlinear ICA has attracted more attention due to its ability to capture the nonlinear relationships within the data [13]. However, the complexity of finding robust regularized nonlinear transformations makes it difficult to use in many situations. Furthermore, nonlinear ICA has the following shortcomings: (i) It requires more training data, especially in higher dimensional situations; (ii) our framework needs to be revised according to the form of the nonlinear ICA. In this case, local linear ICA (which will be used in this paper) presents a good trade-off [14].

Local ICA uses piece-wise linear ICA transformations to approximate nonlinear ICA. In practice, a clustering algorithm is applied first to divide the data into segments. We assume the data within segments have the linear relationship, thus we can use the previously proposed framework to estimate MI within each segment. According to the principle of information addition, total MI equals to the summation of the MIs of each segment. In this way, we can extend our previous framework easily to local ICA-based nonlinear MI estimation. The system block diagram of local ICA-based MI estimation is shown in Fig. 1. In the following sections, we will introduce the theoretic feasibility of local ICA for MI estimation. To verify the proposed approach, we also present experimental results on synthetic and EEG data sets.

II. Theoretical Background

Consider a group of nonlinearly distributed, d-dimensional feature vectors: $\mathbf{x}=[x_1,x_2,...,x_d]^T$. We first apply a suitable clustering algorithm to segment the data into n partitions: $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, …, $\mathbf{x}^{(n)}$. We assume within each partition $\mathbf{x}^{(i)}$, the data is d dimensional, and distributed in accordance with the linear ICA model. Within each partition, we apply the linear ICA transformation to get feature vectors: $\mathbf{y}^{(1)}$, $\mathbf{y}^{(2)}$, …, $\mathbf{y}^{(n)}$, where $\mathbf{y}^{(i)}=[y_1^{(i)},y_2^{(i)},\ldots,y_d^{(i)}]^T$. Since the linear ICA transformation does not change the mutual information, we have:

$$I_S(\mathbf{x}^{(i)};\mathbf{c}) = I_S(\mathbf{y}^{(i)};\mathbf{c}) \tag{1}$$

where $i = 1\ldots n$, denote the n clusters, $\mathbf{x}^{(i)}$ are the original dependent components, and $\mathbf{y}^{(i)}$ are the independent components.

It is well known that the mutual information can be expressed in terms of conditional entropy as follows:

$$I_S(\mathbf{y}^{(i)};c) = H_S(\mathbf{y}^{(i)}) - \sum_c p_c H_S(\mathbf{y}^{(i)} \mid c) \tag{2}$$

where c is the class label, p_c are the prior class probabilities.

If the components of the random vector $\mathbf{y}^{(i)}$ are independent, the joint and joint-conditional entropies become the sum of marginal and marginal-conditional entropies:[1]

$$\begin{aligned} I_S(y_1^{(i)}, y_2^{(i)}, \ldots, y_d^{(i)};\mathbf{c}) \approx & \sum_d H_S(y_d^{(i)}) \\ & - \sum_c P_c \sum_d H_S(y_d^{(i)} \mid \mathbf{c}) \end{aligned} \tag{3}$$

It is easy to show that the total MI between $\mathbf{x}$ and $\boldsymbol{c}$ equals to the summation of the MI within each segmented data group:

$$I_S(\mathbf{x};\mathbf{c}) = \sum_i I_S(\mathbf{x}^{(i)};\mathbf{c}) \tag{4}$$

Combining (1)-(4), we can get:

$$\begin{aligned} I_S(x_1, x_2, \ldots, x_d;\mathbf{c}) \approx \sum_i \Bigg\{ & \sum_d H_S(y_d^{(i)}) \\ & - \sum_c P_c \sum_d H_S(y_d^{(i)} \mid \mathbf{c}) \Bigg\} \end{aligned} \tag{5}$$

In principle, the proposed local ICA for MI estimation contains three parts: clustering algorithm, linear ICA algorithm, and one-dimensional entropy

[1] Here we assume that the same ICA transformation achieves independence in the overall conditional distributions simultaneously.

estimator. For each component, one can use any established method. In this paper, we use K-means clustering, GED-based ICA transformations, and sample-spacing entropy estimators.

K-means clustering: The K-means method is defined as to minimize the over all distance of the data to the center of K clusters:

$$J = \sum_{k} \sum_{x^{(i)} \in S_i} \left\| x^{(i)} - m_i \right\|^2 \tag{6}$$

where $\mathbf{m}_i$ is the center of each cluster. This algorithm first selects K arbitrary cluster centers, and then calculates the distance between all data points to these clusters center respectively. By grouping the nearest data to the each center in a particular group, we get a segmentation of the data set. Because we select the cluster center arbitrarily, this segmentation may not minimize eq (6). Replacing the each cluster center with the mean value of each group, we get a new set of cluster centers. The process above is repeated until J converges to its minimum value.

ICA Using Generalized Eigenvalue Decomposition: The square linear ICA problem is expressed in (7), where **X** is the $n \times N$ observation matrix, **A** is the $n \times n$ mixing matrix, and **S** is the $n \times N$ independent source matrix.

$$\mathbf{X} = \mathbf{AS} \tag{7}$$

Each column of **X** and **S** represents one sample of data. Many effective and efficient algorithms based on a variety of assumptions including maximization of non-Gaussianity and minimization of mutual information exist to solve the ICA problem [11,15,16]. Those utilizing fourth order cumulants could be compactly formulated in the form of a generalized eigen-decomposition problem that gives the ICA solution in an analytical form [10].

According to this formulation, one possible assumption set that leads to an ICA solution utilizes the higher order statistics (specifically fourth-order cumulants). Under this set of assumptions, the separation matrix **W** is the solution to the following generalized eigendecomposition problem:

$$\mathbf{R_x W} = \mathbf{Q_x W \Lambda} \tag{8}$$

where $\mathbf{R_x}$ is the covariance matrix and $\mathbf{Q_x}$ is the cumulant matrix estimated using sample averages: $\mathbf{Q_x}$=E[$\mathbf{x^T x x x^T}$]-$\mathbf{R_x}$tr($\mathbf{R_x}$)-E[$\mathbf{x x^T}$]E[$\mathbf{x x^T}$]-$\mathbf{R_x R_x}$. Given the estimates for these matrices, the ICA solution can be easily determined using efficient generalized eigendecomposition algorithms (or using the *eig* command in Matlab).

Estimating MI Using Sample-Spacings: There exist many entropy estimators in the literature for single-dimensional variables. Here, we use an estimator based on sample-spacings [11], which stems from order statistics. This estimator is selected because of its consistency, rapid asymptotic convergence, and simplicity.

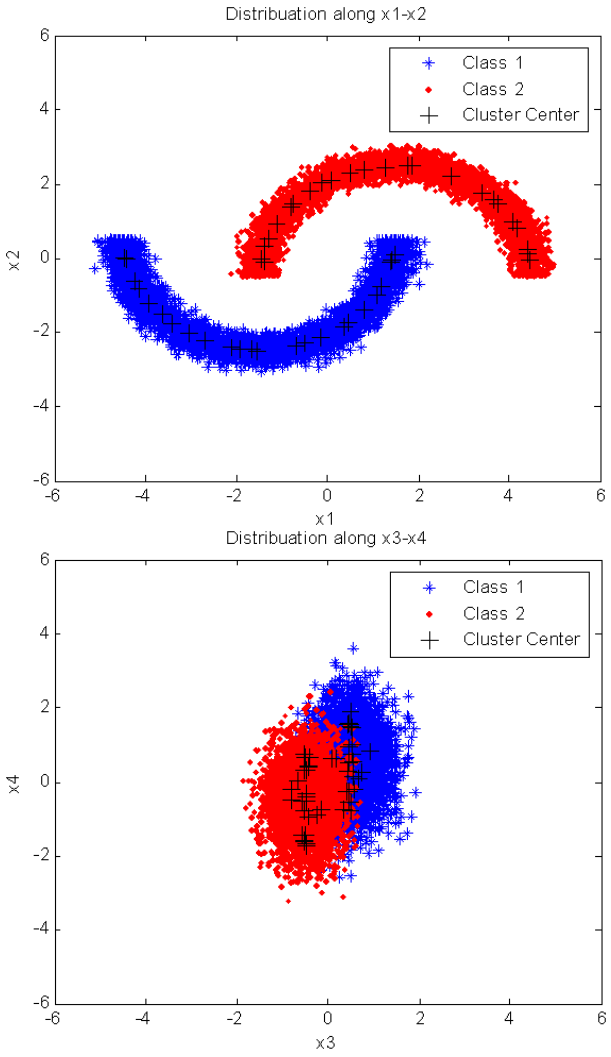

Figure 2. Synthetic dataset. Top: distribution of x_1 and x_2, bottom: distribution of x_3 and x_4.

Consider a one dimensional random variable Y. Given a set of iid samples of Y $\{y_1,\ldots,y_N\}$, first these samples are sorted in increasing order such that $y_{(1)} \leq \ldots \leq y_{(N)}$. The m-spacing entropy estimator is given by:

$$\hat{H}(Y) = \frac{1}{N-m} \sum_{i=1}^{N-m} \log \frac{(N+1)(y_{(i+m)} - y_{(i)})}{m} \tag{9}$$

The selection of the parameter m is determined by a bias-variance trade-off and typically $m = \sqrt{N}$. In general, for asymptotic consistency the sequence $m(N)$ should satisfy

$$\lim_{N \to \infty} m(N) = \infty \qquad \lim_{N \to \infty} m(N)/N = 0 \tag{10}$$

III. Experiments and Results

Synthetic Dataset: In order to illustrate the feasibility and the performance of the proposed local ICA for MI estimation approach, we apply it to a synthetic dataset. This dataset consists of four dimensional feature vectors: x_i (i=1,…,4), where x_1 and x_2 are nonlinearly related (Fig 2 top), x_3 and x_4 are Gaussian distributed with different mean and variance (Fig. 2 bottom). There are two classes in this dataset (represented as blue and red in Fig. 2). These two classes are separable in the x_1 and x_2 plane, but overlapping in the x_3 and x_4 plane. It is clear that this dataset can be well classified by only using x_1 and x_2, while x_3 and x_4 provides redundant and insufficient information for classification. From the Fig. 2 we can see that x_2 has less overlap compared with x_1, while x_3 has less overlap than x_4. So ideally, the feature ranking in descending order of importance in terms of classification rate should be x_2, x_1, x_3, x_4. In our experiments, we choose the sample size as 10000, and segment data into 50 partitions. The '+' in Fig.2 represents the cluster centers. In the top figure we can see these centers distribute evenly along the curve of each classes, segmenting the nonlinear data into piecewise linear groups. The choice of the number of centers K is very critical. By choosing large K, the linear relationship within each segment becomes stronger. However, a large K will cause diminishing sample size within some partitions, which will cause inaccurate ICA transformation estimates and entropy estimates.

By applying the proposed method on this synthesis dataset, we get the feature ranking result: x_2, x_1, x_3, x_4, which is consistent with our expectation.

EEG data classification: To compare the performance of local ICA for MI estimation with the previous linear ICA approximation for MI estimation, we apply both approaches on an EEG-based brain computer interface (BCI) dataset. The EEG data is collected as part of an augmented cognition project, in which the estimated cognitive state is used to assess the mental load of the subject in order to modify the interaction of the subject with a computer system with the goal of increasing user performance. During data collection, two subjects are required to execute different levels of mental tasks, which are classified as high workload and low workload. EEG data is collected using a BioSemi Active Two system using a 31 channel (AF3, AF4, C3, C4, CP1, CP2, CP5, CP6, Cz, F3, F4, F7, F8, FC1, FC2, FC5, FC6, Fp1, Fp2, Fz, O1, O2, Oz, P3, P4, P7, P8, PO3, PO4, Pz) EEG cap and eye electrodes. Vertical and horizontal eye movements and blinks were recorded with electrodes below and lateral to the left eye. EEG is sampled and recorded at 256Hz from 31 channels.

EEG signals are pre-processed to remove eye blinks using an adaptive linear filter based on the Widrow-Hoff training rule (LMS) [17]. Information from the VEOGLB ocular reference channel was used as the noise reference source for the adaptive ocular filter. DC drifts were removed using high pass filters (0.5Hz cut-off). A band pass filter (between 2Hz and 50Hz) was also employed, as this interval is generally associated with cognitive activity. The power spectral density (PSD) of the EEG signals, estimated using the Welch method [18] with 50%-overlapping 1-second windows, is integrated over 5 frequency bands: 4-8Hz (theta), 8-12Hz (alpha), 12-16Hz (low beta), 16-30Hz (high beta), 30-44Hz (gamma). These bands, sampled every 0.25 seconds, are used as the basic features for cognitive classification.

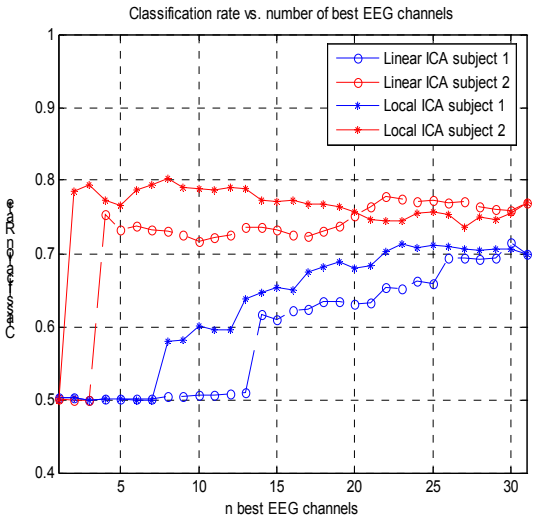

Figure 3 EEG channel ranking in terms of classification rate for two subjects by linear ICA and local linear ICA.

The novelty in this application is that the subjects are freely moving around in contrast to the typical brain-computer interface (BCI) experimental setups where the subjects are in a strictly controlled setting. The assessment of cognitive state in ambulatory subjects is particularly difficult, since the movements introduce strong artifacts irrelevant to the mental task/load. Furthermore, the features extracted from EEG data exhibit strong nonlinearity. In this case, feature selection by local ICA becomes important due to its abilities to precisely keep the useful information and eliminate the irrelevant information for classification.

To test the performance of local ICA for MI estimation in feature selection, we apply the EEG data into the classification system, which contains four parts: preprocessing, feature extraction and selection, classification, and postprocessing. Preprocessing is used to filter out noise and remove the artifacts as mentioned above. Feature extraction and selection generates features from the clean EEG signal, and selects useful EEG channels (each channel contains 5 frequency

bands) using the proposed method. Consider we have around 2500 data samples for each subject, and the dimension of feature is 155 (31 EEG channels, 5 frequency band each), we use K=4, which means to segment data into 4 groups. This is a roughly approximation. However, we can not choose K to be a larger number because on average we only have around 600 data samples for each group with 155 dimensions. For classification, the K-Nearest-Neighbor (KNN) classifier is utilized. The postprocessing uses the assumption that the variations in cognitive state for a given continuous task will be slowly varying in time. A median filter operating on a window of 2-second decisions recently generated by the classifier is used to eliminate a portion of erroneous decisions made by the classification system.

The EEG channel selection results evaluated by classification rate are shown in Fig. 3. The red lines represent the classification performance with different number of optimal EEG channels for subject 1, while the blue lines are for subject 2. As a comparison, we also illustrate the performance using linear ICA for EEG channel ranking on both subjects. The solid line with stars illustrates the classification results for local ICA, while, the dashed line with circles illustrates the classification results for linear ICA. From Fig. 3 we can see that the ability to find an optimum subset for local ICA is superior to linear ICA: for subject 1, we get better classification performance from 7 optimal EEG channels; for subject 2, we get better performance from 1 optimal EEG.

To compare feature ranking/selection results for linear ICA and local ICA more clearly, we list the EEG channel ranking in descend order in terms of contribution to classification for both subjects in the Table.

Table. EEG channel ranking (descending order) in terms of contribution to classification rate for subject 1 and subject 2 with linear ICA and local ICA methods.

Subject	Method	EEG channel ranking
Sub 1	Linear ICA	FC2, AF3, CPZ, FP1, CP5, CP1, C4, CP6, P3, CP2, F4, F3, PO4, O2, P4, O1, PZ, P8, FCZ, FC1, FC6, AF4, FC5, FZ, P7, F8, CZ, FP2, F7, PO3, OZ
	Local ICA	FC2, AF3, CPZ, AF4, FC5, F7, CZ, O2, F3, F4, FC6, C4, F8, P3, FP2, CP6, P8, PZ, P7, FZ, FC1, OZ, PO3, FCZ, FP1, CP2, CP1, P4, CP5, PO4, O1
Sub 2	Linear ICA	FC1, CP1, CZ, O1, C4, F3, FCZ, FC2, FZ, CP2, AF3, FP1, CP6, F4, P3, CPZ, CP5, AF4, FC6, P7, PO4, OZ, PZ, PO3, P4, F8, FC5, O2, F7, FP2, P8
	Local ICA	CP1, O1, FP1, CZ, 'FC1, P8, PO4, FP2, FCZ, P7, F4, P3, P4, PO3, CP6, FC6, CPZ, FC5, AF4, FZ, F3, CP5, F7, F8, AF3, CP2, C4, PZ, FC2, O2, OZ

IV. Conclusions

In this paper, we presented a local ICA approach for MI estimation in feature selection. This work is an extension to our previous proposal of using linear ICA transformations plus sample-spacing entropy estimators for MI estimation. The local ICA approach combines the clustering algorithm of choice (K-means in this paper), the cumulant-based analytical solution for linear ICA transformations, and a computationally efficient mutual information estimator, by taking into account the fact that minimization of Bayes classification error can be approximately achieved by maximizing the mutual information between the features and the class labels. Local ICA is used to approximate nonlinear ICA piecewise linearly for data whose distribution is nonlinearly generated. This yields increased performance and accuracy compared with linear ICA transformations. In theory, nonlinear ICA can perfectly transform dependent components into independent feature vectors; however, the complexity and requirement for large amount of data samples limit its applications. On the other hand, local ICA is relatively simple compared with nonlinear ICA and requires less data samples, which is preferred in most of the applications. We must mention that in order to estimate MI precisely, both nonlinear and local ICA need a large data set, while local ICA leave much free space for a trade-off between accuracy and complexity.

Experiments using synthetic and real (EEG) data demonstrate the utility, feasibility and the effectiveness of the proposed technique. In comparison with the previous linear ICA approach, it is observed (as expected) that local linear ICA yields better performance than linear ICA.

Acknowledgements

This work was supported by DARPA under contract DAAD-16-03-C-0054. The EEG data was collected at the Human-Centered Systems Lab., Honeywell, Minneapolis, Minnesota.

References

[1] T. Lan, D. Erdogmus, A. Adami, M. Pavel, "Feature Selection by Independent Component Analysis and Mutual Information Maximization in EEG Signal Classification," accepted to IJCNN 2005, Montreal, Canada, 2005.

[2] R.M. Fano, *Transmission of Information: A Statistical Theory of Communications*, Wiley, New York, 1961.

[3] M.E. Hellman, J. Raviv, “Probability of Error, Equivocation and the Chernoff Bound,” IEEE Transactions on Information Theory, vol. 16, pp. 368-372, 1970.
[4] K. Torkkola, “Feature Extraction by Non-Parametric Mutual Information Maximization,” Journal of Machine Learning Research, vol. 3, pp. 1415-1438, 2003.
[5] R. Battiti, “Using Mutual Information for Selecting Features in Supervised Neural Networks learning,” IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537-550, 1994.
[6] A. Ai-ani, M. Deriche, “An Optimal Feature Selection Technique Using the Concept of Mutual Information,” Proceedings of ISSPA, pp. 477-480, 2001.
[7] N. Kwak, C-H. Choi, “Input Feature Selection for Classification Problems,” IEEE Transactions on Neural Networks, vol. 13, no. 1, pp. 143-159, 2002.
[9] H.H. Yang, J. Moody, “Feature Selection Based on Joint Mutual Information,” in Advances in Intelligent Data Analysis and Computational Intelligent Methods and Application, 1999.
[9] H.H. Yang, J. Moody, “Data Visualization and Feature Selection: New Algorithms for Nongaussian Data,” Advances in NIPS, pp. 687-693, 2000.
[10] L. Parra, P. Sajda, “Blind Source Separation via Generalized Eigenvalue Decomposition,” Journal of Machine Learning Research, vol. 4, pp. 1261-1269, 2003.
[11] E.G. Learned-Miller, J.W. Fisher III, “ICA Using Spacings Estimates of Entropy,”.Journal of Machine Learning Research, vol. 4, pp. 1271-1295, 2003.
[12] Deniz Erdogmus, Andre Adami, Michael Pavel, Tian Lan, Santosh Mathan, Stephen Whitlow, Michael Dorneich, Cognitive State Estimation Based on EEG for Augmented Cognition. Proceedings of the 2nd International IEEE EMBS Conference on Neural Engineering, Arlington, Virginia, Mar. 16-19, 2005.
[13] C. Jutten, J. Karhunen, “Advances in Nonlinear Blind Source Separation,” Proc. of the 4th Int. Symp. on Independent Component Analysis and Blind signal separation (ICA2003), Nara, Japan, April 1-4, 2003, pp. 245-256.
[14] J. Karhunen, S. Malaroiu, “Locally Linear Independent Component Analysis,” Proc. of the Int. Joint Conf. on Neural Networks (IJCNN’99), July, 1999, Washington DC., USA, pp. 882-887
[15] K.E. Hild II, D. Erdogmus, J.C. Principe, “Blind Source Separation Using Renyi's Mutual Information”, IEEE Signal Processing Letters, vol. 8, no. 6, pp. 174-176, 2001.
[16] A. Hyvärinen, E. Oja, “A Fast Fixed Point Algorithm for Independent Component Analysis”, Neural Computation, vol. 9, no. 7, pp. 1483-1492, 1997.
[17] B. Widrow and M. E. Hoff, "Adaptive switching circuits," in IRE WESCON Convention Record, 1960, pp. 96-104.
[18] P. Welch, “The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short Modified Periodograms”, IEEE Transactions on Audio and Electroacoustics, vol. 15, no. 2, pp. 70-73, 1967.