Journal of VLSI Signal Processing 2007

DOI: 10.1007/s11265-006-0026-5

# Maximally Informative Feature and Sensor Selection in Pattern Recognition Using Local and Global Independent Component Analysis

TIAN LAN AND DENIZ ERDOGMUS
*Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR, USA*

*Received: 24 April 2006; Revised: 8 November 2006; Accepted: 22 December 2006*

**Abstract.** In pattern recognition, a suitable criterion for feature selection is the mutual information (MI) between feature vectors and class labels. Estimating MI in high dimensional feature spaces is problematic in terms of computation load and accuracy. We propose an independent component analysis based MI estimation (ICA-MI) methodology for feature selection. This simplifies the high dimensional MI estimation problem into multiple one-dimensional MI estimation problems. Nonlinear ICA transformation is achieved using piecewise local linear approximation on partitions in the feature space, which allows the exploitation of the additivity property of entropy and the simplicity of linear ICA algorithms. Number of partitions controls the tradeoff between more accurate approximation of the nonlinear data topology and small-sample statistical variations in estimation. We test the ICA-MI feature selection framework on synthetic, UCI repository, and EEG activity classification problems. Experiments demonstrate, as expected, that the selection of the number of partitions for local linear ICA is highly problem dependent and must be carried out properly through cross validation. When this is done properly, the proposed ICA-MI feature selection framework yields feature ranking results that are comparable to the *optimal* probability of error based feature ranking and selection strategy at a much lower computational load.

**Keywords:** feature selection, sensor selection, mutual information, independent component analysis

## 1. Introduction

Feature selection and dimensionality reduction is an important problem for pattern recognition and many other applications. In the pattern recognition context, feature selection and dimensionality reduction can exploit the salient features and eliminate the irrelevant features. This results in increased robustness and improved generalization performance of the classification system. Dimensionality reduction can be achieved by subspace projection or its special case feature selection. In subspace projection, the original features are projected linearly or nonlinearly to a low dimensional space, which preserves the desirable characteristics of the data. There are many existing subspace projection methods, such as PCA, ICA and LDA [1–5]. However, the projections that PCA and ICA seek are unsupervised and not necessarily related to the classification performance. LDA overcomes this shortcoming by finding the projections that maximize class separability assuming class-Gaussianity. Torkkola [6] proposed an approach using a quadratic divergence measure to find an optimal transformation that maximizes the MI between features and class labels. This approach, being dependent on Parzen density estimation, is inefficient for subspace projections to high dimensionalities due to the joint density estimation requirement.

Lan et al. developed a subspace projection framework, which applies linear ICA transformation and mutual information maximization for dimensionality reduction in EEG signal classification [7]. This method exhibits several advantages, such as computationally efficiency and flexibility. However, the linearity assumption in ICA limits its applications.

Although subspace projections can effectively remove redundant features, the relationship between the projected features and the original features becomes vague. In some applications, such as multi-sensor array target detection and dense-array EEG signal processing, a given system can only collect and process signals from a certain number of sensors in real-time, due to the limitation of bandwidth and computation capacity. In these particular cases, feature (or sensor) selection is more suitable, which selects a subset from the original feature space. It is widely accepted that some classification algorithms, such as decision tree, multi-layer perceptron neural networks have inherent ability to focus on relevant features and ignore irrelevant ones [8]. In general, feature selection is achieved by a feature ranking procedure. Feature selection methods can be divided into wrapper and filter approaches. Wrapper approach uses classification accuracy as the criterion coupled with a specific classifier; it requires re-training the classifier for different combinations of feature sets; hence, it is slow and inflexible. Filter approach, on the other hand, ranks and selects features by optimizing some criteria independent of the classifier, and is more flexible and suitable for adaptive learning.

In the filter approach, it is important to optimize a criterion that is relevant to Bayes risk, which is typically measured by the probability of error. A suitable criterion is the MI between the selected features and the class labels, motivated by lower and upper bounds in information theory that relate this quantity to probability of error [9, 10]. As opposed to linear and second-order statistics such as correlation and covariance, MI measures nonlinear dependencies between a set of random variables taking into account higher order statistical structures existing in the data.

Many feature selection methods have been developed in the past years [11–13]. Guyon & Elisseeff also reviewed several approaches used in the context of machine learning [14]. By extending our previous work in dimensionality reduction [7], we propose an ICA-MI framework for feature selection. We exploit the fact that an invertible linear transformation does not change the MI, and assume that linear ICA transformation yields independent features (globally or locally). So we can conveniently estimate the MI between feature vectors and class labels by directly summing the MI between each independent projected feature vector and class labels. In cases where the linearity assumption does not hold, we use local linear ICA to approximate nonlinear ICA and extend the ICA-MI framework.

## 2. ICA-MI Feature Selection Framework

The MI based method for feature selection is motivated by lower and upper bounds in information theory [9, 10]. 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 the feature vectors and class labels. Specifically, Hellman & Raviv showed that an upper bound on Bayes error is given by $(H(C)\text{-}I(\mathbf{x},C))/2$, where $H(C)$ is the Shannon entropy of the priori probabilities of the classes and $I(\mathbf{x},C)$ is the Shannon MI between the continuous-valued feature vector and the discrete-valued class label. Maximizing this MI reduces the upper bound as well as Fano's lower bound, therefore, forces the probability of error to decrease.

### 2.1. Mutual Information

In feature selection, we are interested in the MI between the continuous-valued feature vector $\mathbf{x}$ and the discrete-valued class labels $C$. Shannon MI between $\mathbf{x}$ and $C$ is defined in terms of the entropies of the overall data and the individual classes as

$$I(\mathbf{x};C) = H(\mathbf{x}) - \sum_c p_c H(\mathbf{x}/c) \tag{1}$$

where $p_c$ are the prior class probabilities. The entropy is given by

$$\begin{aligned} H(\mathbf{x}) &= -\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} \\ H(\mathbf{x}|C) &= -\int p(\mathbf{x}|C) \log p(\mathbf{x}|C) d\mathbf{x} \end{aligned} \tag{2}$$

where $p(\mathbf{x}|C)$ are the class conditional distributions and the overall data distribution is

$$p(\mathbf{x}) = \sum_c p_c p(\mathbf{x}|c) \tag{3}$$

### 2.2. Mutual Information Decomposition

Directly estimating MI between feature vectors **x** and class label $C$ requires estimating entropy and conditional entropy, which is difficult for high dimensional data. If the components of **x** are mutually independent, the high dimensional joint entropy can be obtained by the summation of marginal entropies. However, in real word applications, **x** usually has mutually dependent components. Assume that there exists a linear transformation that transforms **x** to **y** such that components of **y** are mutually independent.[1] Since an invertible transformation does not change mutual information, we have $I(\mathbf{x};C)=I(\mathbf{y};C)$, and

$$I(\mathbf{y};C) = \sum_{i=1}^{n} I(y_i;C) \tag{4}$$

where $I(y_i;C) = H(y_i) - \sum_c p_c H(y_i|c)$, and $y_i$ is the $i$th component of features. So we convert a high dimensional MI estimation problem to the ICA transformation and marginal MI estimation problem.

### 2.3. ICA-MI Feature Selection

Given a high dimensional feature vector **x**, our goal is to find the best $m$ dimensional subset of features (in terms of maximum MI with $C$). This is a combinatorial search problem, and often $m$ is not defined a priori. An alternative strategy is to rank the features and pick the top $m$ features from this ranking. Given previously ranked $d$-1 features $x_{(1)},\ldots, x_{(d-1)}$ the $d$th feature is the one that maximizes the joint MI: $I(x_{(1)},\ldots, x_{(d-1)},x_{(?)};C)$. The joint mutual information takes into account any redundancies in the new feature with the previously ranked $d-1$ features. This ranking procedure requires the repeated evaluation of $d$-dimensional MI values. Assume the linear assumption holds, so the MI can be estimated in the following procedure: (1) Apply linear ICA on $d$ dimensional overall data **x** consisting of the previously ranked $d-1$ features and the current candidate feature; repeat this using data only from class $C$: $\mathbf{x}^c$, get independent features **y**, and $\mathbf{y}^c$, the transformation matrixes are **W** and $\mathbf{W}^c$; (2) Estimate entropy and conditional entropy of $H(\mathbf{y})$ and $H(\mathbf{y}|C)$; (3) Estimate mutual information $I(\mathbf{x};C)$. The framework of linear ICA-MI for feature selection is illustrated in Fig. 1a.

If the linearity assumption does not hold, a nonlinear ICA transformation is desirable to achieve independent ***y***. Nonlinear ICA requires more data samples and is computationally intense. Furthermore, if the transformation is not invertible, the mutual information changes after transformation. So it is not possible to estimate MI using the proposed framework. Karhunen et al. proposed a local linear ICA algorithm that uses piecewise linear ICA to approximate nonlinear ICA [15]. The idea of local linear ICA is: first segment the data into $p$ partitions, then assume that the linearity assumption holds in each partition, and apply linear ICA within each partition. It is easy to extend linear ICA-MI to local linear ICA-MI (see Fig. 1b). Also, when the number of partitions equals 1, local linear ICA reduces to

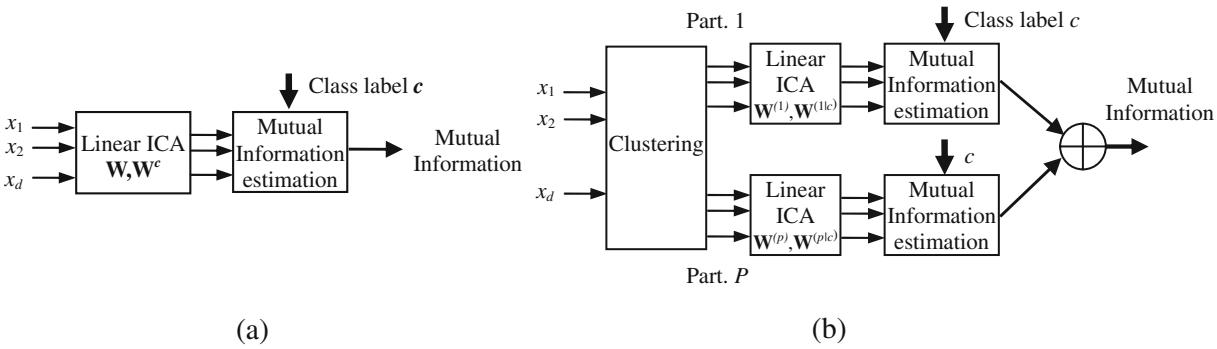


*Figure 1.* Block Diagram of ICA-MI framework for feature selection. **a** *Linear* ICA; **b** *Local linear* ICA.

linear ICA. We can formulate both cases together and the algorithm is described as follows: First apply a suitable clustering/quantization algorithm to segment the data into $p$ partitions: $\boldsymbol{x}^{(1)}$, $\boldsymbol{x}^{(2)}$, ..., $\boldsymbol{x}^{(p)}$; assume that within each partition $\boldsymbol{x}^{(i)}$, the data is $d$ dimensional (at the $d$th step of the ranking procedure, this vector is comprised of previously ranked $d-1$ features and the candidate feature from the unranked ones), and distributed in accordance with the linear ICA model; apply the linear ICA transformation on each partition $C$+1 times (where $C$ is the number of classes. We apply ICA on overall data $\boldsymbol{x}^{(i)}$ and data from class C, $\boldsymbol{x}^{(i|c)}$) to get transformed feature vectors for each partition: $\boldsymbol{y}^{(i|c)}$, which is transformed with class specific linear ICA matrix $\mathbf{W}^{(i|c)}$ and $\mathbf{y}^{(i)}$, which is transformed with the overall partition ICA matrix $\mathbf{W}^{(i)}$, where $C$ denotes class labels. As a result of the linear ICA transformations, we have:

$$\begin{aligned} H\left(x^{(i)}\right) &= H\left(y^{(i)}\right) - \log\left|\mathbf{W}^{(i)}\right| \\ H\left(x^{(i|c)}\right) &= H\left(y^{(i|c)}\right) - \log\left|\mathbf{W}^{(i|c)}\right| \end{aligned} \tag{5}$$

where $i$=1,...,$p$. $H(\mathbf{x}^{(i)})$ is the entropy for cluster *(i)*, $H(\mathbf{x}^{(i/c)})$ is the conditional entropy for cluster *(i)* in class $C$. If linear ICA works perfectly, then the joint entropies of $\mathbf{y}^{(i|c)}$ and $\mathbf{y}^{(i)}$ reduce to the sum of marginal entropies. However, this is not guaranteed, therefore, the residual mutual information will remain as an estimation bias. In practice, we have an imperfect ICA solution and

$$\begin{aligned} H\left(\mathbf{x}^{(i)}\right) &= \sum_{l=1}^{d} H\left(y_l^{(i)}\right) - \log\left|\mathbf{W}^{(i)}\right| - I\left(\mathbf{y}^{(i)}\right) \\ H\left(\mathbf{x}^{(i|c)}\right) &= \sum_{l=1}^{d} H\left(y_l^{(i|c)}\right) - \log\left|\mathbf{W}^{(i|c)}\right| - I\left(\mathbf{y}^{(i|c)}\right) \end{aligned} \tag{6}$$

Mutual information satisfies the following additivity property for any partition ($q_i$ denoting the probability mass of the corresponding partition):

$$I(\mathbf{x}; C) = \sum_{i=1}^{p} q_i I\left(\mathbf{x}^{(i)}, c\right) \tag{7}$$

The mutual information within each partition can be expressed as a linear combination of entropy values as follows:

$$I\left(\mathbf{x}^{(i)}, C\right) = H\left(\mathbf{x}^{(i)}\right) - \sum_{c} p_{ic} H\left(\mathbf{x}^{(i)}|c\right) \tag{8}$$

where $p_{ic}$ denotes the probability mass of class C in partition $i$. Substituting Eq. (6) in Eq. (8)

$$\begin{aligned} I\left(\mathbf{x}^{(i)}, C\right) = &\left(\sum_{l=1}^{d} H\left(y_l^{(i)}\right) - \sum_{c} p_{ic} \sum_{l=1}^{d} H\left(y_l^{(i|c)}\right)\right) \\ &- \left(\log\left|\mathbf{W}^{i}\right| - \sum_{c} p_{ic} \log\left|\mathbf{W}^{i|c}\right|\right) \\ &- \left(I\left(\mathbf{y}^{(i)}\right) - \sum_{c} p_{ic} I\left(\mathbf{y}^{(i|c)}\right)\right) \end{aligned} \tag{9}$$

The last parenthesis in Eq. (9) shows the estimation bias one makes when estimating the MI within each partition if it is assumed that the local linear ICA solution in that partition achieved perfect separation. Over all partitions, the total estimation bias (estimated MI minus the actual MI) is averaged as follows:

$$Bias = \sum_{i=1}^{p} q_i \left(I\left(y^{(i)}\right) - \sum_{c} p_{ic} I\left(y^{(i|c)}\right)\right) \tag{10}$$

Note that as the number of partitions approach infinity asymptotically, one could utilize a grid partitioning structure within which the probability distributions would be uniform, thus local linear ICA would achieve perfect separation within each infinitesimal hypercube. However, in practice, one cannot utilize infinitely many partitions given a finite number of samples. Note that the analysis above also holds for the case where linear ICA is employed directly on the whole dataset without any partitions.

The decomposition of mutual information into overlapping segments (cover rather than partition) has been previously studied by Szummer and Jaakkola in the context of model regularization in the presence of unlabeled data and semi-supervised learning [16]. The partition approach we propose here is along the same direction of reasoning, that is the cumulative relevant information of a feature

vector can be decomposed to local regions in the vector space; however, while Szummer and Jaakkola are interested in emphasizing discriminative and dense regions in the data for density fitting, we are interested in estimating the total useful information in a feature vector.

## 3. Linear ICA and Local Linear ICA

In our experience, the bias of MI estimation does not seem to influence the feature selection results significantly. Eq. (10) indicates that this bias depends on the performance of the ICA assumption as well as the particular algorithm used in obtaining separation solutions. The linear ICA signal model is as follows [4, 5, 17]: there are $n$ independent sources $\mathbf{s}(t)=[s_1(t), ..., s_n(t)]^{\mathrm{T}}$, and $n$ observations $\mathbf{x}(t)=[x_1(t), ..., x_n(t)]^{T}$, where $t$ is the sample index. The observations are mixed versions of the sources by a full rank matrix $\mathbf{A}$:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{11}$$

The goal of ICA is to find a separation matrix $\mathbf{W}$ such that

$$\mathbf{y}(t) = \mathbf{W}x(t) \tag{12}$$

where $\mathbf{y}(t)=[y_1(t), ..., y_n(t)]^{\mathrm{T}}$ are mutually independent. Given arbitrary $n$-dimensional observations, one can always find infinitely many nonlinear transformations $\mathbf{y}=\mathbf{f}(\mathbf{x})$ that result in independent components $\mathbf{y}=\{y_1,...,y_n\}$ [18]. In practice, however, especially in situations involving small datasets, finding a robust nonlinear ICA is difficult without a priori information about the data distribution. A convenient and suitable way to solve the nonlinear ICA problem is to approximate it in a piecewise linear fashion using local linear ICA.

The main principle of local linear ICA is that to segment the data into $p$ non-overlapping partitions. In theory, for infinitesimal partitions the linearity assumption always holds within partitions; however, small sample size prevents reliable linear ICA estimates. Therefore, the tradeoff between the number of partitions and samples per partition must be considered in the bias-variance framework. Cross-validation can be used to determine the proper number of partitions.

## 4. Materials and Methods

We focus on the local linear ICA as a general case. The data is partitioned using the K-means clustering algorithm [19]. Linear ICA is solved in each partition using generalized eigendecomposition of 2nd and 4th order cumulant matrices. One-dimensional entropies are estimated using the sample spacing approach.

### 4.1. K-means Clustering

The K-means algorithm tries to minimize the average squared distance of the data to the centers of clusters:

$$J = \sum_{i} \sum_{\mathbf{x}^{(i)} \in S_i} \left\| \mathbf{x}^{(i)} - \mathbf{m}_i \right\|^2 \tag{13}$$

where $\mathbf{m}_i$ is the center of each cluster. This algorithm first selects $K$ random cluster centers, and then calculates the distance between all data points to these clusters center, respectively. Samples are assigned to the cluster corresponding to the nearest center and then cluster centers are updated to the average of the assigned samples. The process is repeated until $J$ converges to its minimum value (local minimum).

### 4.2. ICA Using Generalized Eigendecomposition of Cumulant Matrices

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 [20–22]. Those utilizing fourth order cumulant could be compactly formulated in the form of a generalized eigen-decomposition problem that gives the ICA solution in an analytical form [23].

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

$$\mathbf{R}_\mathbf{x}\mathbf{W} = \mathbf{Q}_\mathbf{x}\mathbf{W}\Lambda \tag{14}$$

where $\mathbf{R}_\mathbf{x}$ is the covariance matrix and $\mathbf{Q}_\mathbf{x}$ is the cumulant matrix estimated using sample averages: $\mathbf{Q}_x = \mathrm{E}\left[\mathbf{x}^\mathbf{T}\mathbf{x}\mathbf{x}\mathbf{x}^\mathbf{H}\right] - \mathbf{R}_\mathbf{x}\mathrm{tr}(\mathbf{R}_\mathbf{x}) - \mathrm{E}\left[\mathbf{x}\mathbf{x}^\mathbf{T}\right]\mathrm{E}\left[\mathbf{x}\mathbf{x}^\mathbf{H}\right] - \mathbf{R}_\mathbf{x}\mathbf{R}_\mathbf{x}$.

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).

### 4.3. Entropy Estimator

We employ an estimator based on sample spacing [20], which stems from order statistics. This estimator is selected because of its consistency, rapid asymptotic convergence, and simplicity. Consider a one dimensional random variable $Y$. Given a set of iid samples of $Y$ $\{y_1,...,y_N\}$, first these samples are sorted in increasing order such that $y_{(1)}\leq...\leq y_{(N)}$. The $m$-spacing entropy estimator is given by:

$$\widehat{H}(Y) = \frac{1}{N-m}\sum_{i=1}^{N-m}\log\frac{(N+1)\left(y_{(i+m)}-y_{(i)}\right)}{m} \quad (15)$$

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

$$\lim_{N\to\infty} m(N) = \infty \lim_{N\to\infty} m(N)/N = 0 \quad (16)$$

### 4.4. Feature Ranking Algorithm (Quasi-greedy Search)

We use an incremental strategy to rank and select features instead of exhaustive combinatorial search. Given $n$-dimensional feature vector $\mathbf{x}$ with corresponding class label variable C, the incremental ranking proceeds as follows:

1. Let *Unranked Feature Set* (UFS) be $\{x_1,...,x_n\}$ Estimate the MI $I(x_j,C)$ between each feature $x_j$ ($j$=1,...,$n$) and class label $C$. Find the feature with maximum MI, label it as $x_{(1)}$, initialize *Ranked Feature Set* (RFS) to$\{x_{(1)}\}$ and remove the feature corresponding to $x_{(1)}$ from the UFS.
2. For $d$ from 2 to $n$ perform the following: Let *Candidate Set i* (CS$i$) be the union of RFS and $x_i$, evaluate the MI $I$(CS$i$,$C$) between the features in the candidate set and the class labels for every $x_i$ in UFS. Label the feature $x_i$ with the highest $I$(CS$i$,$C$) as $x_{(d)}$, redefine RFS as the union of RFS and $x_{(d)}$, and remove the corresponding feature from UFS.

## 5. Experimental Results

### 5.1. Experiments on a Synthetic Dataset

In order to illustrate the difference between linear ICA and local ICA for MI based feature selection, we apply both approaches on a synthetic dataset. This dataset consists of four features: $x_i$ ($i$=1,...,4), where $x_1$ and $x_2$ are nonlinearly related (Fig. 2-left), $x_3$ and $x_4$ are independent from the first two features and are linearly correlated Gaussian-distributed with different mean and variance (Fig. 2-right). There are two classes in this dataset (represented as different grayscale levels in print). 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 only using $x_1$ and $x_2$, while $x_3$ and $x_4$ provides redundant and insufficient information for perfect classification. From Fig. 1 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 1,000 and use 20 partitions.[2] The '+' in Fig. 2 represents the partition centers. We also apply linear ICA without any partitioning. We repeat the above experiment for 100 Monte Carlo runs. The linear ICA approach finds the ranking to be $x_2$, $x_1$, $x_4$, $x_3$, while the local linear ICA approach with 20 partitions finds the expected *correct* ranking.

### 5.2. Experiments on UCI Dataset

***5.2.1. Iris Data.*** In this experiment, we apply linear and local linear ICA (with 2 partitions) approaches to the ranking of the features for the Iris dataset from the UCI database [24]. Due to the small sample size, 10 Monte Carlo rankings with randomly selected training (used for ranking) and test sets are utilized, each consisting of 50% of the available samples. For each ranked subset, a Gaussian Mixture Model (GMM) based Bayesian classifier is employed. The frequency of rankings and classification accuracy are shown in Table 1 and Fig. 3. Since both methods agree on $x_4$ as the top one, pairwise scatter plots of this feature with the remaining features are shown in

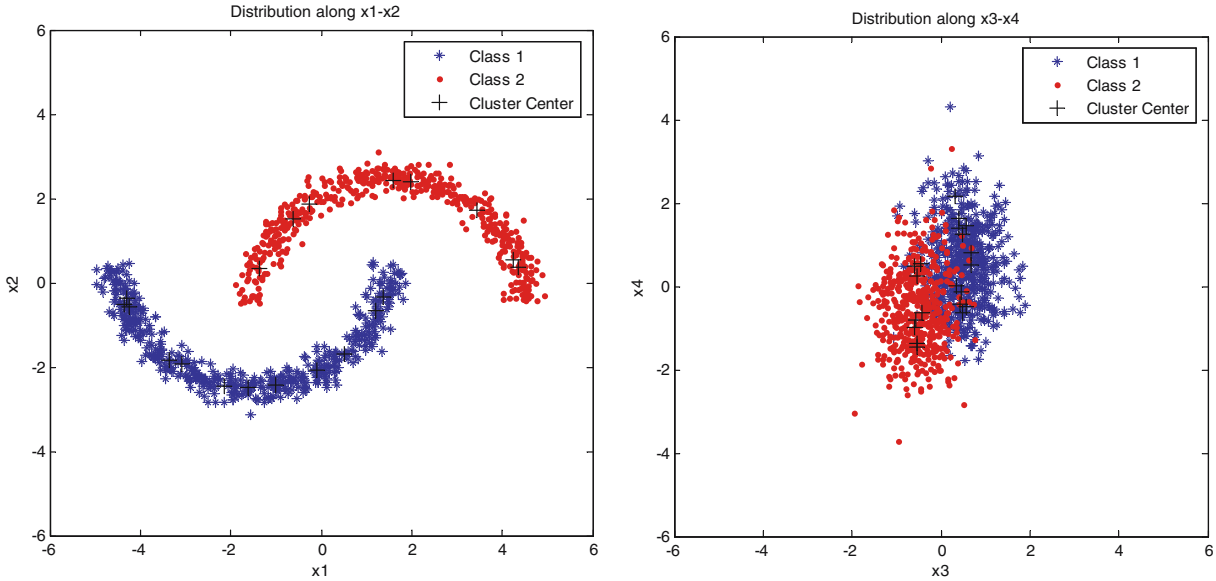


*Figure 2.* Four-dimensional Synthetic dataset and corresponding cluster centers. *Left*: distribution of $x_1$ and $x_2$; *Right*: distribution of $x_3$ and $x_4$.

Fig. 4 for visual comparison. $x_3$ seems to yield a more compact class distribution, while $x_1$ or $x_2$ and $x_4$ seem to have less overlapping samples. Still, it is difficult to judge and we rely on the GMM performances on the testing set for the final comparison. The classification accuracy in Fig. 3 shows that local linear ICA yields more accurate feature ranking than linear ICA in Iris data.

***5.2.2. Wisconsin Breast Cancer Data.*** We apply linear ICA and local linear ICA for feature selection on Wisconsin breast cancer dataset, which has higher dimensionality than the previous two case studies. Local linear ICA approach uses 2 partitions (for lack of sufficient data per partition otherwise) and the Monte Carlo ranking approach is employed as 5.2.1. The ranking and classification accuracy are shown in Table 2 and Fig. 5. Local linear ICA also exhibit better performance than linear ICA. Consider the number of data samples and the dimensions: if we partition the data into more segments, the performance degrades due to the lack of data for reliable linear ICA transformation within each partition.

*Table 1.* Feature ranking frequencies on the Iris dataset.

| Methods | Ranking indices | | | | |
|---|---|---|---|---|---|
| Linear ICA | 4 | 3 | 2 | 1 | (10) |
| Local linear ICA | 4 | 1 | 2 | 3 | (5) |
| | 4 | 2 | 3 | 1 | (3) |
| | 4 | 2 | 1 | 3 | (2) |

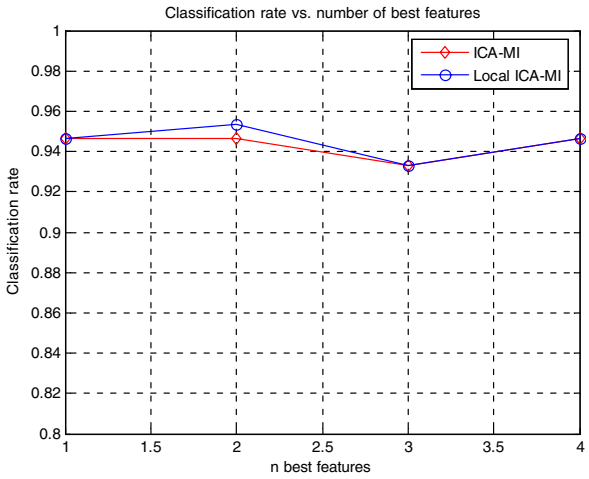


*Figure 3.* Classification accuracy for Iris data by linear ICA-MI and Local linear ICA-MI methods. The classification accuracy is the average over 10 Monte Carlo simulations.

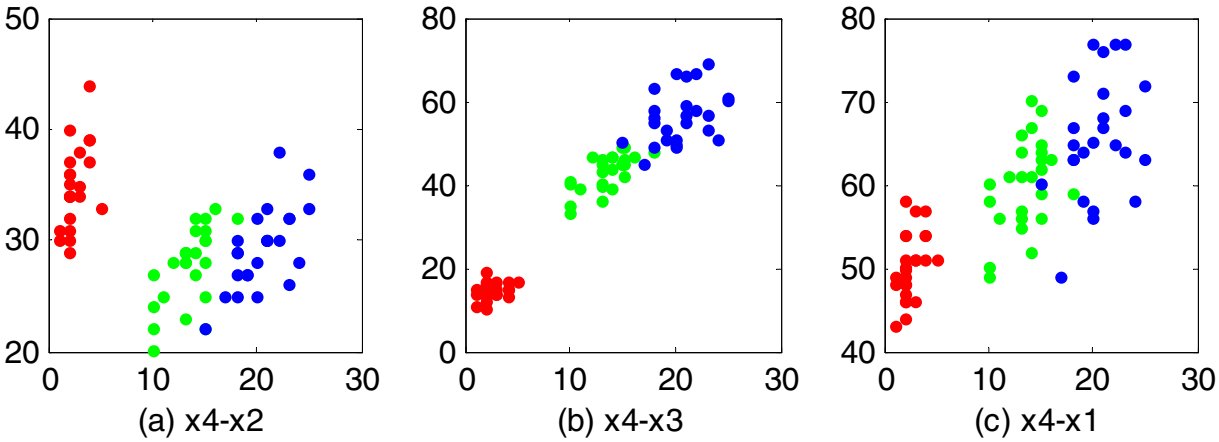


*Figure 4.* Combinational distribution of 2 feature vectors of Iris dataset. *Left*: distribution of $x_4$ and $x_2$; *Middle*: distribution of $x_4$ and $x_3$; *Right*: distribution of $x_4$ and $x_1$.

### 5.3. *Experiments on BCI Competition Dataset*

In this experiment, we apply both methods on Brain Computer Interfaces Competition III dataset V [25]. This dataset contains human brain EEG data from 3 subjects during 4 non-feedback sessions. The subject sat in a chair, relaxed arms resting on their legs and executed one of the three tasks: imagination of left hand movements; imagination of right hand movements; generation of words beginning with the same random letter. The EEG data were collected during the sessions. The data of all 4 sessions of a given subject were collected on the same day, each lasting 4 min with 5–10 min breaks in between. We want to classify one of the three tasks from the EEG data. The raw EEG data contains 32 channels at 512 Hz sampling rate. The raw EEG potentials were first spatially filtered by means of a surface Laplacian. Then, every 62.5 ms, the power spectral density (PSD) in the band 8–30 Hz was estimated over the last second of data with a frequency resolution of 2 Hz for the eight centro-parietal channels C3, Cz, C4, CP1, CP2, P3, Pz, and P4. As a result, an EEG sample is a 96-dimensional vector (8 channels $\times$ 12 frequency components).

We apply both ICA feature selection methods on this dataset, and then use Support Vector Machine (SVM) to classify them. For the SVM, we use Chang & Lin's library toolbox [26]. Based on the experiment results, we select the parameter of SVM as: penalty parameter $c$=10, and kernel size $g$=10. We mix the first three sessions as training set and use the

*Table 2.* Feature ranking results on Wisconsin Breast Cancer dataset for different ICA-MI methods in 10 Monte Carlo simulations.

| Methods | Ranking indices | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Linear ICA | 3 | 2 | 9 | 4 | 5 | 6 | 7 | 8 | 1 | (9) |
| | 3 | 2 | 9 | 4 | 5 | 8 | 7 | 6 | 1 | (1) |
| Local linear ICA | 3 | 1 | 2 | 4 | 5 | 6 | 7 | 8 | 9 | (4) |
| | 3 | 4 | 6 | 8 | 7 | 1 | 9 | 2 | 5 | (3) |
| | 3 | 1 | 4 | 5 | 9 | 6 | 8 | 2 | 7 | (3) |

The frequency of different ranking of 10 Monte Carlo simulations are shown inside the bracket.

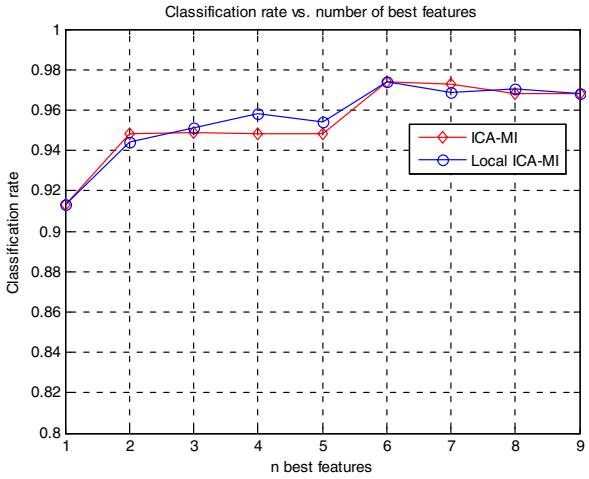


*Figure 5.* Classification accuracy for Wisconsin Breast Cancer data by different ICA-MI methods. The classification accuracy is the average over 10 Monte Carlo simulations.

fourth session as testing set. For the local linear ICA, we choose the cluster number $K$ as 2, 5 and 10, respectively.

As a comparison, we also apply Bayes error based feature ranking using SVM. Two of the first three sessions are used as training, and the other session is used as validation. Rotating the training and validation sets, and applying 3-fold cross-validation, we get the Bayes error based feature ranking using the incremental searching strategy. Then we test the error based ranking using SVM on the fourth session. On average, the error based ranking is optimal for the selected classifier and the given search strategy. The feature ranking results for 3 subjects are shown in Tables 3, 4, and 5. The classification results based on the best features are shown in Figs. 6, 7, and 8. The tables and figures only show the first 30 features.

The experiment results illustrate that: (1) Error based feature ranking yields the best performance on average; (2) For subject 1 and 2, linear ICA has better performance than local linear ICA, and the performance is close to error based feature ranking; this also indicates the linear relationship among feature vectors for subject 1 and 2; (3) For subject 3, local linear ICA has better performance than linear ICA, which indicate nonlinear relationship of the data; (4) For $K$=2, 5, 10, the feature ranking results does not change drastically.

### 5.4. *Experiments on AugCog Dataset*

The comparison is repeated for the EEG classification problem encountered in the context of Augmented Cognition [27]. The EEG activity is measured to estimate the cognitive state in order to assess the mental load of the subject for the purpose of modifying the computer/system information interface. The goal is to increase the task performance of the subject with closed loop cognitive interface control. During data collection, two subjects are asked to execute different mental tasks (playing an action video game at different difficulty levels), which are classified as high workload and low workload. EEG data is collected using a BioSemi Active Two system using 31 channels (AF3, AF4, 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, CPz, FCz) EEG cap and eye electrodes. Vertical and horizontal eye movements

*Table 3.* Feature ranking results on BCI competition III dataset V subject 1.

| Method | Feature ranking indices |
|---|---|
| Error based | 27 38 2 25 86 49 13 35 4 3 77 20 81 62 36 10 18 48 26 32 31 80 55 46 84 82 60 78 21 83 |
| Linear ICA | 38 3 26 27 58 2 4 87 39 85 46 68 91 7 35 33 31 32 52 92 37 72 70 63 75 18 8 79 90 41 |
| Local Linear ICA $K$=2 | 38 2 32 14 3 4 39 80 58 66 18 83 91 35 77 50 7 73 21 19 9 8 12 5 84 54 88 15 40 79 |
| Local Linear ICA $K$=5 | 38 3 2 14 77 32 8 4 94 49 7 6 9 90 19 18 54 21 12 37 84 36 33 56 39 70 67 34 60 20 |
| Local Linear ICA $K$=10 | 38 2 50 3 90 35 12 4 5 8 9 92 1 53 7 18 88 80 48 33 22 96 69 73 77 43 36 41 15 39 |

*Table 4.* Feature Ranking results on BCI competition III dataset V subject 2.

| Method | Feature ranking indices |
|---|---|
| Error based | 2 26 95 6 90 5 11 31 23 38 33 77 82 96 81 48 55 60 84 70 24 76 93 94 32 37 12 41 57 47 |
| Linear ICA | 26 2 45 13 85 34 74 75 58 27 21 94 89 73 68 1 4 50 84 7 72 96 46 29 71 60 22 63 20 11 |
| Local Linear ICA $K$=2 | 26 2 36 84 27 17 33 12 29 6 75 58 92 80 72 53 56 67 44 48 22 77 59 11 82 34 30 21 20 55 |
| Local Linear ICA $K$=5 | 26 2 33 83 11 93 58 34 42 46 29 16 95 94 19 36 38 54 7 81 50 79 48 82 10 43 80 90 8 75 |
| Local Linear ICA $K$=10 | 26 2 17 74 15 31 11 60 43 67 50 89 47 82 54 33 58 36 20 24 30 48 51 34 38 9 91 41 70 78 |

*Table 5.* Feature Ranking results on BCI competition III dataset V subject 3.

| Method | Feature ranking indices |
|---|---|
| Error based | 3 4 21 39 28 74 25 52 48 67 94 76 45 9 17 55 88 43 78 83 84 57 51 11 5 24 79 19 31 80 |
| Linear ICA | 3 74 93 31 90 30 95 4 39 65 44 34 1 5 35 12 24 10 58 51 8 84 11 37 53 69 43 21 23 81 |
| Local Linear ICA $K$=2 | 3 45 39 4 92 30 8 90 28 94 7 22 20 46 9 47 48 82 23 69 1 78 12 68 11 93 96 19 79 55 |
| Local Linear ICA $K$=5 | 3 9 4 1 80 92 29 35 31 36 52 28 11 66 76 96 39 30 93 95 32 46 12 90 23 78 54 20 65 72 |
| Local Linear ICA $K$=10 | 3 4 72 39 9 45 93 38 6 22 1 96 23 79 66 7 53 48 69 77 12 10 18 36 59 78 83 20 61 33 |

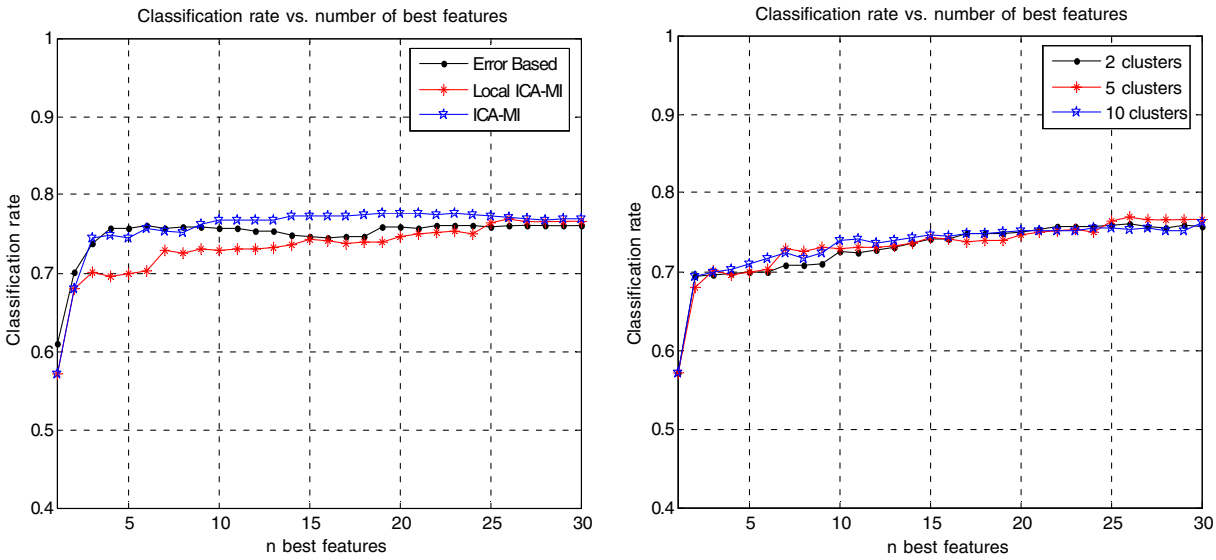


*Figure 6.* Classification accuracy for BCI competition III dataset V subject 1 by different feature ranking method. The cluster number $K$=5 for the local ICA-MI method in *left* figure.

and blinks were recorded with electrodes below and lateral to the left eye. EEG is sampled and recorded at 256 Hz.

EEG signals are pre-processed to remove eye blinks using an adaptive linear filter based on the Widrow-Hoff training rule (LMS) [28]. 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 (0.5 Hz cut-off). A band pass filter (between 2 and 50 Hz) 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 [29] with 75%-overlapping 1 s windows, is integrated over 5 frequency bands: 4–8 Hz (theta), 8–12 Hz (alpha), 12–16 Hz

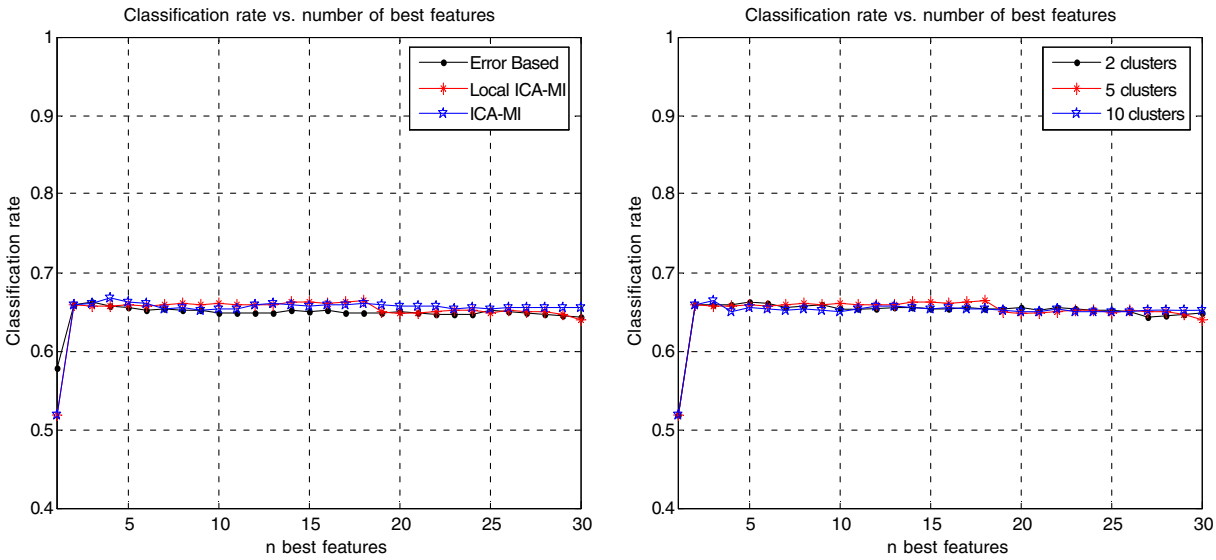


*Figure 7.* Classification accuracy for BCI competition III dataset V subject 2 by different feature ranking method. The cluster number $K$=5 for the local ICA-MI method in *left* figure.

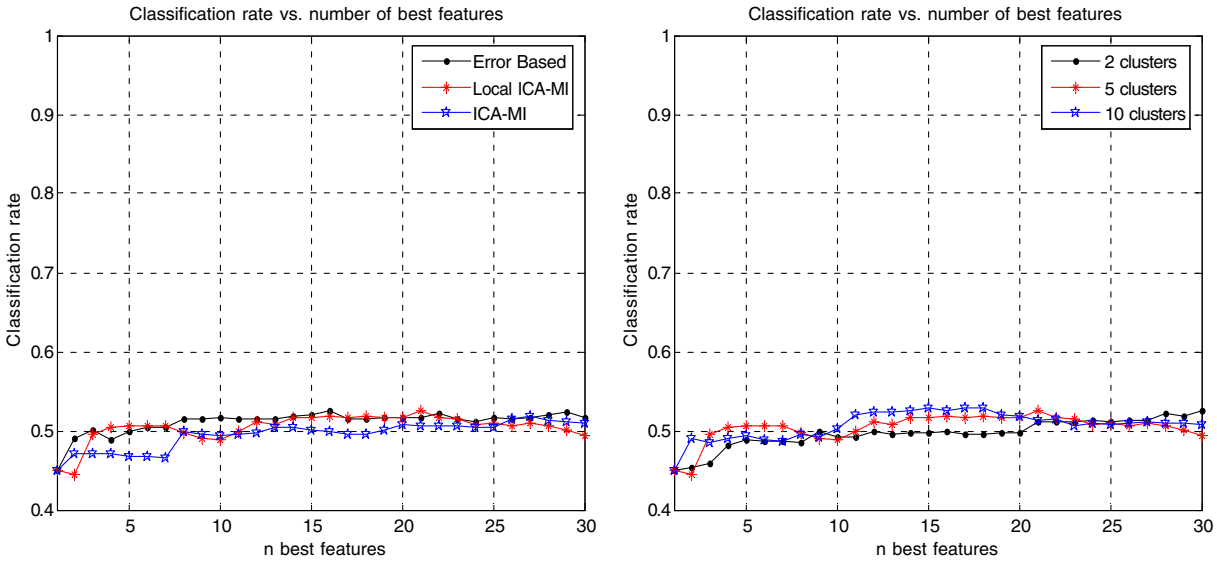


*Figure 8.* Classification accuracy for BCI competition III dataset V subject 3 by different feature ranking method. The cluster number $K$=5 for the local ICA-MI method in *left* figure.

(low beta), 16–30 Hz (high beta), 30–44 Hz (gamma). These bands, sampled every 0.25 s, are used as the basic features for cognitive classification.

The novelty in this application is that the subjects are freely moving 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.

To test the performance of local ICA for MI estimation in feature selection, the EEG data is processed by a classification system that 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. We have approximately 2,500 data samples for each subject, and the number of features is 155 (31 EEG channels, 5 frequency band each). We use $K$=4 to have an average of 600 samples per partition. For classification, a K-Nearest-Neighbor (KNN) classifier with 11 neighbors is utilized (GMM-Bayes and SVM classifiers performed poorly on this dataset). The postprocessing uses the assumption that the cognitive state for a given continuous task will vary slowly in time. A median filter operating on a window of 2.25 s eliminates a portion of outlier from the decisions recently generated by the classifier.

The EEG channel selection results evaluated by correct classification rate are shown in Fig. 9 (when a channel is selected/discarded all five features

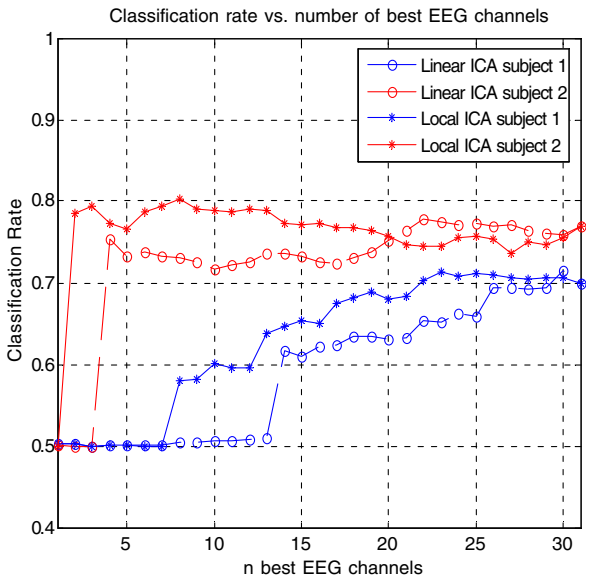


*Figure 9.* EEG channel ranking in terms of classification rate for two subjects by *linear* ICA and *local linear* ICA.

*Table 6.* 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 |

associated with the channel are selected/discarded). 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 for both subjects. We observe that local ICA outperforms linear ICA in both subjects. EEG channels ranked according to both methods are shown in Table 6.

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 6.

## 6. Discussion

In this paper, we propose a local linear ICA—maximum mutual information framework for feature selection in pattern recognition. As a special case of this framework, the linear ICA-MI approach is included for a single data partition. This framework contains 3 components: (1) clustering algorithm to partition the feature space; (2) linear ICA transformation to project the data within a partition to an independent coordinate frame; (3) marginal entropy estimator.

The proposed framework has the following advantages: (1) using mutual information between feature vectors and class labels consider the data structure together with class separability; (2) this is a filter approach, so it is flexible and computationally efficient; (3) it works well in high dimensions; (4) it is a general framework; any component can be replaced by suitable counterparts.

We applied both linear ICA-MI and local linear ICA-MI with different number of partitions on different datasets. In the synthetic dataset, because we construct the data with a nonlinear structure, local linear ICA-MI obtains more accurate feature selection results. In UCI datasets (Iris and Wisconsin Breast Cancer), local linear ICA also exhibits better performance than linear ICA. In BCI competition III dataset V, linear ICA has better performance in the datasets from two subjects and local linear ICA has better performance for the third subject. This indicates that the selection between linear and local linear ICA is data and application dependent. In the AugCog dataset, for both subjects local linear ICA outperforms linear ICA in EEG channel selection.

An issue is how to select the number of partitions. In most applications, we have only limited data. Therefore, the number of partitions controls the trade-off between bias and variance. When the number of partitions is very large, the linearity assumption holds well within each partition, but there is not have enough data to apply linear ICA within partitions. On the contrary, when the number of partitions is very small, local linear ICA converge to linear ICA, which is not proper for nonlinear data. A good method to select this parameter is cross-validation, which helps us identify whether the available data supports a nonlinear model.

## Acknowledgement

This research was partially supported by DARPR through DAAD-16-03-C-0054, HM1582-05-C-0046, and by NSF through ECS-0524835. The data used in the AugCog experiment was colleted at the Honeywell Human-Centered System Laboratory. The authors thank Misha Pavel for valuable discussions.

## Notes

1. We assume there exists a linear transformation that decomposes the overall data $\mathbf{x}$ to $\mathbf{y}$, as well as decomposes the data for each class $\mathbf{x}^c$ to $\mathbf{y}^c$. This assumption does not hold in most of cases, we usually apply different transformation to overall data and data from different classes, respectively. We will discuss this in detail in Section 2.3.
2. The optimal number of partitions can be achieved by M-fold cross-validation process.

## References


1. E. Oja, *Subspace Methods of Pattern Recognition*, Wiley, New York, 1983.
2. P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*, Prentice Hall, London, 1982.
3. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed., Academic, New York, 1990.
4. R. Everson and S. Roberts, "Independent Component Analysis: A Flexible Nonlinearity and Decorrelating Manifold Approach," *Neural Comput*., vol. 11, no. 8, 2003, pp. 1957–1983.
5. A. Hyvärinen, E. Oja, P. Hoyer, and J. Hurri, "Image Feature Extraction by Sparse Coding and Independent Component Analysis," *Proc. of ICPR'98*, 1998, pp. 1268–1273.
6. K. Torkkola, "Feature Extraction by Non-parametric Mutual Information Maximization," *J. Mach. Learn. Res*., vol. 3, 2003, pp. 1415–1438.
7. T. Lan, D. Erdogmus, A. Adami, and M. Pavel, "Feature Selection by Independent Component Analysis and Mutual Information Maximization in EEG Signal Classification," *Proc. of IJCNN'05*, Montreal, 2005, pp. 3011–3016.
8. W. Duch, T. Wieczorek, J. Biesiada, and M. Blachnik, "Comparison of Feature Ranking Methods Based on Information Entropy," *Proc. of International Joint Conference on Neural Networks (IJCNN)*, IEEE Press, Budapest, 2004, pp. 1415–1420.
9. R. M. Fano, *Transmission of Information: A Statistical Theory of Communications*, Wiley, New York, 1961.
10. M. E. Hellman and J. Raviv, "Probability of Error, Equivocation and the Chernoff Bound," *IEEE Trans. Inf. Theory*, vol. 16, 1970, pp. 368–372.
11. R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Training," *IEEE Trans. Neural Netw*., vol. 5, no. 4, 1994, pp. 537–550.
12. K. Kira and L. Rendell, "The Feature Selection Problem: Traditional Methods and a New Algorithm," in *Proc. of the Tenth National Conference on Artificial Intelligence (AAAI-92)*, Menlo Park, 1992, pp. 129–134.
13. G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant Features and the Subset Selection Problem," in *Proc. of the 11th International Conference on Machine Learning*, San Mateo, 1994, pp. 121–129.
14. I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *J. Mach. Learn. Res.* (Special Issue on Variable and Feature Selection), 2003.
15. J. Karhunen, S. Malaroiu, and M. Ilmoniemi, "Local linear independent component analysis based on clustering," *Int. J. Neural Syst*., vol. 10, no. 6, 2000, pp. 439–451.
16. M. Szummer and T. Jaakkola, "Information Regularization with Partially Labeled Data," Advances in NIPS 15, 2002.
17. S. Haykin (Ed.), *Unsupervised Adaptive Filtering, Vol. I: Blind Source Separation*, 2000, Wiley, York.
18. A. Hyvarinen and P. Pajunen, "Nonlinear Independent Component Analysis: Existence and Uniqueness Results," *Neural Netw*., vol. 12, no. 3, 1999, pp. 429–439.
19. S. Haykin, *Neural Networks—A Comprehensive Foundation*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1998.
20. O.Vasicek, "A Test for Normality Based on Sample Entropy," *J. R. Stat. Soc., Ser. B*, vol. 38, no. 1, 1976, pp. 54–59.
21. K. E. Hild II, D. Erdogmus, and J. C. Principe, "Blind Source Separation Using Renyi's Mutual Information," *IEEE Signal Process. Lett*., vol. 8, no. 6, 2001, pp. 174–176.
22. A. Hyvärinen and E. Oja, "A Fast Fixed Point Algorithm for Independent Component Analysis," *Neural Comput*., vol. 9, no. 7, 1997, pp. 1483–1492.
23. L. Parra and P. Sajda, "Blind Source Separation via Generalized Eigenvalue Decomposition," *J. Mach. Learn. Res*., vol. 4, 2003, pp. 1261–1269.
24. http://www.ics.uci.edu/~mlearn/MLRepository.html, UCI Machine Learning Repository.
25. http://ida.first.fraunhofer.de/projects/bci/competition_iii/#data_set_v, BCI Competition III.
26. C. C. Chang and C. C. Lin, LIBSVM: A Library for Support Vector Machines, 2001, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
27. S. Mathan, N. Mazaeva, S. Whitlow, A. Adami, D. Erdogmus, T. Lan, and M. Pavel, "Sensor-based Cognitive State Assessment in a Mobile Environment," *Proc. of AUGCOG'05 (jointly with HCII'05)*, 2005, Las Vegas, Nevada.
28. B. Widrow and M. E. Hoff, "Adaptive Switching Circuits," in *IRE WESCON Convention Record*, 1960, pp. 96–104.
29. 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 Trans. Audio Electroacoust*., vol. 15, no. 2, 1967, pp. 70–73.


**Tian Lan** received both his B.S. and M.S. degrees in Electronics Engineering from Beijing University of Aeronautics and Astronautics in 1997 and 2000 respectively. He is currently a Ph.D.

candidate and research assistant in Electrical Engineering at OGI School of Science & Engineering, Oregon Health & Science University. His research work focuses on signal processing, adaptive system, machine learning, and information theory, and their applications in biomedical signal processing, including EEG, MRI, and brain computer interfaces.

**Deniz Erdogmus** received his B.S. degree in Electrical & Electronics Engineering (EEE), and his B.S. degree in Mathematics both in 1997, and his M.S. degree in EEE in 1999 from the Middle East Technical University, Turkey. He received his Ph.D. in Electrical & Computer Engineering from the University of Florida (UF) in 2002. He worked as a research engineer at TUBITAK-SAGE, Turkey from 1997 to 1999, focusing on the design of navigation, guidance, and flight control systems. He was also a research assistant and a postdoctoral research associate at UF from 1999 to 2004, concentrating on signal processing, adaptive systems, machine learning, and information theory, specifically with applications in biomedical engineering including brain machine interfaces. Currently, he is holding an Assistant Professor position jointly at the Computer Science and Electrical Engineering Department and the Biomedical Engineering Department of the Oregon Health and Science University. His research focuses on information theoretic adaptive signal processing and its applications to biomedical signal processing problems. Dr. Erdogmus has over 50 articles in international scientific journals and numerous conference papers and book chapters. He has also served as associate editor and guest editor for various journals, participated in various conference organization and scientific committees, and he is a member of TBP, HKN, IEEE, IEE, INNS, and EURASIP.