Extensions of Independent Component Analysis Mixture Models for classification and prediction of EEG signals

Gonzalo Safont[1], Addisson Salazar[1], Alberto Rodriguez[2], Luis Vergara[1]
Correspondence author: asalazar@dcom.upv.es

[1] *Institute of Telecommunications and Multimedia Applications*
Universitat Politècnica de València
Camino de Vera s/n, 46022, Valencia, Spain

[2] *Signal Processing Group, Communications Engineering Dpt.*
Universidad Miguel Hernández de Elche
Avda. Universidad S/N, 03202 Elche, Spain

Abstract

This paper presents two applications of Independent Component Analysis Mixture Modeling (ICAMM) for the classification and prediction of data. The first one of these extensions is Sequential ICAMM (SICAMM), an ICAMM structure that takes into account the sequential dependence in the feature record. This algorithm can be used to classify input observations in a given set of mutually-exclusive classes. The performance of SICAMM is tested with simulations and compared against that of the base ICAMM algorithm and of a Dynamic Bayesian Network (DBN). All three methods are also used to classify real electroencephalographic (EEG) signals to compute hypnograms, a clinical tool used to help in the diagnosis of sleep disorders. The second extension of ICAMM is PREDICAMM, an estimation algorithm that makes use of the ICAMM parameters in order to reconstruct missing samples from a set of data. This predictor is used to reconstruct real EEG data from a working memory experiment, and its performance is compared to that of a classical predictor for EEG signals: sphere splines. Prediction performance is measured with four error indicators: signal-to-interference ratio, Kullback-Leibler divergence, correlation, and mean structural similarity index. Both extensions of the base ICAMM algorithm have achieved a higher performance than other methods.

Keywords: ICA mixture model, EEG, prediction, classification, working-memory task.

1. Introduction

Electroencephalographic (EEG) signals are recordings of surface brain electrical activity taken at the scalp, each sensor capturing the combined signal from multiple neurons of the brain [1]. The study of EEG signals is a useful clinical tool because some illnesses, typically seizures and sleep disorders, produce abnormal electric patterns in the electrical activity of the brain that can be identified by an expert. EEG signals are also the subject of much research on brain activity, given the lower hardware cost and higher temporal resolution of EEG when compared to other available techniques.

Some of this research revolves around the processing of EEG signals with Independent Component Analysis (ICA), a blind source separation (BSS) method that is progressively finding more applications for BSS and feature modeling/extraction [2][3][4][5]. ICA algorithms find a linear transformation of the observed signals so as to maximize the statistical independence of the transformed signals (also known as sources). There is extensive literature of ICA in biomedical applications [6][7][8][9][10][11][12]. Some relevant works in EEG - ICA signal processing are the study of developmental differences in the saccadic contingent negative variation [13], EEG and event-related potential (ERP) data [14][15], combination of multiple detectors for EEG biometric identification [16][17], and removal of artifacts in the EEG signal [18].

ICA is extended in the ICA Mixture Model (ICAMM), a method where multiple ICA models are learned and weighed in a probabilistic manner. ICAMM implies conditional independence, since data from the same model are maximally independent, but data can have any amount of inter-class dependencies [3][19][20]. ICAMM has been proposed as a flexible approximation for modeling mixtures of arbitrary probability densities, being able to model local structures without loss of generalization capabilities [21][22][23][24][25]. However, there are few references of ICAMM for EEG signal processing [26][27][28][29].

Some illnesses, typically seizures and sleep disorders, produce abnormal electric patterns in the electrical activity of the brain that can be measured by EEG signals and identified by an expert.

This paper presents two extensions of ICAMM for prediction and classification of EEG signals. The first application is an extension of the basic ICAMM structure to include sequential dependence between classes, also known as SICAMM [26]. SICAMM is applied to the classification of EEG signals from sleep studies, in order to automate the detection of periods of wakefulness in sleeping subjects. This is an automatic detection technique in the line with other techniques and applications where the authors are working currently [30][31].

The second application is the use of ICAMM to estimate missing data, making use of the flexibility of the mixture model: this algorithm is known as PREDICAMM [32][33]. Although there are several applications of ICA as preprocessor in prediction of temporal series (see for instance [34]), the prediction itself has not been done considering an underlying ICA model of the data density. In this work, PREDICAMM is used to estimate missing data from real EEG records and compared against a classical EEG signal predictor, sphere splines.

2. Independent Component Analysis Mixture Models

Independent Component Analysis (ICA) is a blind source separation technique that attempts to find a linear transformation such that the transformed components, also known as sources, are as statistically independent as possible [2]. The standard noiseless ICA model assumes that the observations, $\mathbf{x}(n)=\left[x_1(n),x_2(n),\ldots,x_M(n)\right]^{\mathrm{T}}$, are composed by linear mixtures of random variables that are mutually independent, the sources $\mathbf{s}(n)=\left[s_1(n),s_2(n),\ldots,s_L(n)\right]^{\mathrm{T}}$. That is, for

$$\mathbf{x}(n)=\mathbf{A}\cdot\mathbf{s}(n) \tag{1}$$

ICA methods estimate simultaneously the so-called mixing matrix $\mathbf{A}$ and the sources $\mathbf{s}(n)$. The mixing matrix contains the coefficients of the linear transformations from sources to observations: thus, it can be applied to separate each of the sources $s_l(n)=\mathbf{w}_l\cdot\mathbf{x}(n), l=1,\ldots,L$, where $\mathbf{w}_l$ is the l-th row of $\mathbf{W}$. For the sake of simplicity, we will assume the same number of sources and mixtures, i.e., $L=M$. We will further consider that the mixing matrix can be inverted, i.e. that $\mathbf{W}=\mathbf{A}^{-1}$ exists. It bears noting that the ICA problem is well-defined if and only if at most one source is Gaussian, and the rest are non-Gaussian. If they are indeed Gaussian, the results of any ICA method will be similar to those of Principal Component Analysis (PCA).

Notice the lack of a noise term in (1). Standard ICA methods assume no noise in order to obtain tractable and simple enough estimation algorithms, especially for the estimation of noise-free components. Moreover, there are many cases where data cannot be divided into signals and noise in any meaningful way.

This model can be expanded into a mixture model, forming an ICA Mixture Model (ICAMM). ICAMM were first proposed in the framework of pattern recognition, considering that the observed data can be categorized in K mutually exclusive classes, and that each one of these classes can be adequately modeled through ICA. ICAMM was introduced in [19], considering a source model that could switch between super-Gaussian (Laplacian) and sub-Gaussian (bimodal) probability density functions. A more recent generalization of the ICAMM framework was proposed in [20]. This generalization includes non-parametric density estimation and semi-supervised learning; supports the use of any ICA algorithm for parameter updating; and corrects residual dependencies between sources.

The general formulation of ICAMM is:

$$\mathbf{x}(n)=\mathbf{A}_k\cdot\mathbf{s}_k(n)+\mathbf{b}_k,\ k=1,\ldots,K \tag{2}$$

assuming that $\mathbf{x}(n)$ belongs to class k, that is, $C_l(n)=k$; and class k is described by an ICA model with mixing matrix $\mathbf{A}_k$ and bias vector $\mathbf{b}_k$. Essentially, the bias vector determines the location of the cluster and the mixing matrix determines its shape.

The goal of ICAMM methods is to determine the parameters for each one of the classes. Algorithms for learning the ICAMM parameters in supervised or unsupervised frameworks can be found in the references ([19][35][36][37]). In practice, an expert determines the class of each observation and data are split. Then, the parameters for each class are calculated using any standard ICA method.

In most applications, the objective is classifying each observed data vector in one of the available classes. This is accomplished by computing the posterior probability of the classes given the observation, $P(C_k(n)|\mathbf{X}(n))$, and selecting the class with maximum probability. In practice, these posteriors are calculated with the help of Bayes' theorem, as it allows expressing $P(C_k(n)|\mathbf{X}(n))$ in terms of the probability density function of the observations:

$$p(C_k(n)|\mathbf{x}(n))=\frac{p(\mathbf{x}(n)|C_k(n))\cdot P(C_k(n))}{p(\mathbf{x}(n))}=\frac{p(\mathbf{x}(n)|C_k(n))\cdot P(C_k(n))}{\sum_{l=1}^{K}p(\mathbf{x}(n)|C_l(n))\cdot P(C_l(n))} \tag{3}$$

where the mixture model is evident in the denominator of (3). The conditional probability $p(\mathbf{x}(n)|C_k(n))$ can be calculated using the ICAMM parameters in (2),

$$p(\mathbf{x}(n)|C_k(n))=|\det\mathbf{W}_k|\cdot p(\mathbf{s}_k(n)) \tag{4}$$

where $\mathbf{s}_k(n)=\mathbf{W}_k\cdot(\mathbf{x}(n)-\mathbf{b}_k)$. Therefore, if the classifier has been trained, we can compute the posterior probabilities as:

$$p(C_k(n)|\mathbf{x}(n))=\frac{P(C_k(n))\cdot|\det\mathbf{W}_k|\cdot p(\mathbf{s}_k(n))}{\sum_{l=1}^{K}P(C_l(n))\cdot|\det\mathbf{W}_l|\cdot p(\mathbf{s}_l(n))} \quad (5)$$

2.1. Sequential ICAMM

The ICAMM model assumes that the observations are independent from each other, and that the classes are independent as well. In practice, though, both exhibit some degree of dependence in the time domain. In order to capture this dependence in the ICA Mixture Model, the calculation of the posterior probabilities should consider not just the current observation, but also past observations. That is, the algorithm would maximize the conditional probability $P(C_k(n)|\mathbf{X}(n))$, where $\mathbf{X}(n)=[\mathbf{x}(0),\mathbf{x}(1),\ldots,\mathbf{x}(n)]$ is the historic of all observations up to time instant .

This dependence is considered in Sequential ICAMM (SICAMM) [26], which we will explain here. SICAMM assumes that the statistical dependence between two successive time instants can be modeled with a classical Hidden Markov Model (HMM) [38] structure, with each class corresponding to a different hidden state. This assumption implies that the observations $\mathbf{x}(n)$ are conditionally independent given their classes, $C_k(n)$. Thus, $P(C_k(n)|\mathbf{X}(n)$ can be expressed as the product of the conditional probability of the current observation, times the conditional probability of all past observations: $p(\mathbf{X}(n)|C_k(n))=p(\mathbf{x}(n)|C_k(n))\cdot p(\mathbf{X}(n\text{-}1)|C_k(n))$. Using this property and Bayes' theorem:

$$\begin{aligned}p(C_k(n)|\mathbf{X}(n))&=\frac{p(\mathbf{X}(n)|C_k(n))\cdot P(C_k(n))}{\sum_{l=1}^{K}p(\mathbf{X}(n)|C_l(n))\cdot P(C_l(n))}=\\&=\frac{p(\mathbf{x}(n)|C_k(n))\cdot p(\mathbf{X}(n\text{-}1)|C_k(n))\cdot P(C_k(n))}{\sum_{l=1}^{K}p(\mathbf{x}(n)|C_l(n))\cdot p(\mathbf{X}(n\text{-}1)|C_l(n))\cdot P(C_l(n))}=\\&=\frac{p(\mathbf{x}(n)|C_k(n))\cdot p(C_k(n)|\mathbf{X}(n\text{-}1))\cdot p(\mathbf{X}(n\text{-}1))}{\sum_{l=1}^{K}p(\mathbf{x}(n)|C_l(n))\cdot p(C_l(n)|\mathbf{X}(n\text{-}1))\cdot p(\mathbf{X}(n\text{-}1))}=\\&=\frac{|\det\mathbf{W}_k|\cdot p(\mathbf{s}_k)\cdot p(C_k(n)|\mathbf{X}(n\text{-}1))}{\sum_{l=1}^{K}|\det\mathbf{W}_l|\cdot p(\mathbf{s}_l)\cdot p(C_l(n)|\mathbf{X}(n\text{-}1))}\end{aligned} \quad (6)$$

where, considering the HMM model,

$$P(C_k(n)|\mathbf{X}(n\text{-}1))=\sum_{l=1}^{K}P(C_k(n)|C_l(n\text{-}1))\cdot P(C_l(n\text{-}1)|\mathbf{X}(n\text{-}1)) \quad (7)$$

Notice that the above holds true even if there is no sequential class dependence, in which case $P(C_k(n)|P(C_l(n\text{-}1))=P(C_k(n))$ and thus $P(C_k(n)|\mathbf{X}(n\text{-}1))=P(C_k(n))$.

ICAMM is a flexible approximation for modeling mixtures of arbitrary probability densities, being able to model local structures without loss of generalization capabilities.

Thus, $P(C_k(n)/\mathbf{X}(n))$ can be computed from the class transition probabilities $P(C_k(n)|C_l(n\text{-}1))$ and the last estimated available of the class probabilities, $P(C_k(n\text{-}1)|\mathbf{X}(n\text{-}1))$, using (6) and (7).

The SICAMM algorithm is described as follows:
Initialization:

$$\mathbf{X}(0)=\mathbf{x}(0)$$

$$\mathbf{s}_k(0)=\mathbf{W}_k\cdot(\mathbf{x}(0)-\mathbf{b}_k)$$

$$P(C_k(0)|\mathbf{X}(0))=\frac{|\det\mathbf{W}_k|\cdot p(\mathbf{s}_k(0))}{\sum_{l=1}^{K}|\det\mathbf{W}_l|\cdot p(\mathbf{s}_l(0))}$$

For $n=1,\ldots,N$

$$\mathbf{X}(n)=[\mathbf{x}(0),\mathbf{x}(1),\ldots,\mathbf{x}(n)]$$

$$\mathbf{s}_k(n)=\mathbf{W}_k\cdot(\mathbf{x}(n)-\mathbf{b}_k)$$

$$P(C_k(n)|\mathbf{X}(n\text{-}1))=\sum_{l=1}^{K}P(C_k(n)|C_l(n\text{-}1))\cdot P(C_l(n\text{-}1)|\mathbf{X}(n\text{-}1))$$

$$P(C_k(n)|\mathbf{X}(n))=\frac{|\det\mathbf{W}_k|\cdot p(\mathbf{s}_k(0))\cdot P(C_k(n)|\mathbf{X}(n\text{-}1))}{\sum_{l=1}^{K}|\det\mathbf{W}_l|\cdot p(\mathbf{s}_l(0))\cdot P(C_l(n)|\mathbf{X}(n\text{-}1))}$$

The SICAMM parameters are the ICA model for each class, $\mathbf{W}_k$, $\mathbf{b}_k$, $p(\mathbf{s}_k)$, $k=1,\ldots,K$, and the class transition probabilities $P(C_k(n)|C_l(n\text{-}1))$. The estimation of these parameters depends on the amount of supervision available for training data. For supervised estimation, the ICA models can be learnt separately from the class transition probabilities using any of the algorithms proposed for ICAMM. $P(C_k(n)|C_l(n\text{-}1))$ is usually estimated empirically by counting the transitions between class C_l and class C_k in training data. For semi-supervised estimation, though, all the SICAMM parameters have to be estimated simultaneously.

2.2. PREDICAMM

PREDICAMM is an algorithm that uses known data and ICA Mixture Models to estimate missing data. Let us consider data vector $\mathbf{x}(n)$ of size $[M\ \mathrm{x}\ 1]$ which can be modeled through an ICA Mixture Model such as the one shown in (2). The parameters for this model, $\mathbf{W}_k=\mathbf{A}_k^{-1}$, $p(\mathbf{s}_k(n))$, $\mathbf{b}_k$, have to be estimated from training data using any ICAMM method. Assuming that M_{unk} values of vector $\mathbf{x}(n)$ are unknown, known and unknown values can be grouped into two smaller vectors, $\mathbf{y}(n)$ (known values) and $\mathbf{z}(n)$ (unknown values). That is,

$$\mathbf{x}(n)=[\mathbf{y}^T(n),\mathbf{z}^T(n)]^T \quad (8)$$

The goal is to predict $\mathbf{z}(n)$ using the known values, $\mathbf{y}(n)$. This prediction is calculated by maximizing the joint probability density of the observations, $p(\mathbf{y}(n),\mathbf{z}(n))=p(\mathbf{x}(n))$.

Using (4), the probability density function of the data can be expressed as:

$$p(\mathbf{x}(n))=\sum_{k=1}^{K} p(\mathbf{x}(n)|C_k(n))\cdot p(C_k(n))=\sum_{k=1}^{K} |\det \mathbf{W}_k|\cdot p(\mathbf{s}_k(n))\cdot P(C_k(n)) \quad (9)$$

where $\mathbf{s}_k(n)=\mathbf{W}_k\begin{pmatrix}\mathbf{y}(n)\\ \mathbf{z}(n)\end{pmatrix}-\mathbf{b}_k$. The maximization of this cost function is performed by means of a gradient algorithm, usually steepest descent. This algorithm requires the derivative of the cost function:

$$\frac{\delta p(\mathbf{y}(n),\mathbf{z}(n))}{\delta \mathbf{z}(n)}=\sum_{k=1}^{K} |\det \mathbf{W}_k|\cdot \frac{\delta p(\mathbf{s}_k(n))}{\delta \mathbf{z}(n)}\cdot P(C_k(n)) \quad (10)$$

$$\frac{\delta p(\mathbf{s}_k(n))}{\delta \mathbf{z}(n)}=\sum_{m=1}^{M} \frac{\delta p(\mathbf{s}_k(n))}{\delta s_{k,m}(n)}\frac{\delta s_{k,m}(n)}{\delta \mathbf{z}(n)} \quad (11)$$

where $s_{k,m}(n)$ is the m-th source of class k at time instant n. The derivative $\frac{\delta p(\mathbf{s}_k(n))}{\delta \mathbf{z}(n)}$ can be calculated from the mixture model and (8):

$$\mathbf{W}_k\cdot\mathbf{x}=\mathbf{W}_k\begin{pmatrix}\mathbf{y}(n)\\ \mathbf{z}(n)\end{pmatrix}=[\mathbf{Q}_k\ \mathbf{R}_k]\cdot\begin{pmatrix}\mathbf{y}(n)\\ \mathbf{z}(n)\end{pmatrix}=\mathbf{Q}_k\cdot\mathbf{y}(n)+\mathbf{R}_k\cdot\mathbf{z}(n)=\mathbf{Q}_k\cdot\mathbf{y}(n)+\begin{pmatrix}\mathbf{r}_{k,1}^T\mathbf{z}(n)\\ \vdots\\ \mathbf{r}_{k,M}^T\mathbf{z}(n)\end{pmatrix} \quad (12)$$

where $\mathbf{R}_k$ is a matrix composed by the last M_{unk} columns of matrix $\mathbf{W}_k$, and $\mathbf{r}_{k,m}^T$ is the -th row of $\mathbf{R}_k$. Thus, the sources can be expressed as the sum of two terms, one of which is not dependent on $\mathbf{z}(n)$. The partial derivative of the sources with respect to vector $\mathbf{z}(n)$ is thus:

$$\frac{\delta \mathbf{s}_{k,m}(n)}{\delta \mathbf{z}(n)}=\frac{\delta}{\delta \mathbf{z}(n)}(\mathbf{r}_{k,n}^T\mathbf{z}(n))=\mathbf{r}_{k,n} \quad (13)$$

Assuming that the sources are independent and replacing (13) and (11) in (10), we obtain the target function:

$$\frac{\delta p(\mathbf{y}(n),\mathbf{z}(n))}{\delta \mathbf{z}(n)}=\sum_{k=1}^{K} |\det \mathbf{W}_k|\cdot P(C_k(n))\cdot\sum_{m=1}^{M}\frac{\delta p(\mathbf{s}_k(n))}{\delta \mathbf{s}_{k,m}(n)}\cdot\mathbf{r}_{k,n} \quad (14)$$

The estimation of the probability density function of the sources $p(\mathbf{s}_k(n))$, and its derivatives, is not trivial. In this work, it was estimated using a non-parametric (kernel) density estimator.

3. Dynamic Bayesian Networks

Bayesian Networks (BN) are graphical structures that allow us to represent and reason about an uncertain domain [39][40][41][42][43]. A BN is composed by a directed graph, G, and a set of conditional probability distributions, Θ. The graph is composed of nodes and edges: nodes in the graph represent random variables, and edges state direct dependencies between the nodes (variables) they link. In particular, an edge from node χ_i to node χ_j implies direct statistical dependence between variables i and j. In this context, node χ_i is then referred to as a parent of χ_j and, similarly, χ_j is referred to as a child of χ_i. These terms can be extended to define the sets of descendants or ascendants of a given node.

The only restriction of the graph is that no node that can be its own ancestor or its own descendent, in order to simplify the calculation of the joint probability distribution of the network. Such a network is called a directed acyclic graph or "dag".

BN reflect simple conditional independence statements, since each variable is independent of its non-descendants in the graph given the state of its parents. The graph encodes these independence assumptions while Θ contains the parameters $\Theta_{\chi_i|\pi_i}=P_B(\chi_i|\pi_i)$ for each realization of variable χ_i. Then, the BN can be used to define the joint probability distribution of a set of variables χ_i, $i=1,\ldots,n$:

$$P_B(\chi_1,\chi_2,\ldots,\chi_n)=\prod_{i=1}^{n} P_B(\chi_i|\pi_i)=\prod_{i=1}^{n}\Theta_{\chi_i|\pi_i} \quad (15)$$

Dynamic Bayesian Networks (DBN) are an extension of BN developed to deal with stochastic processes. In this case, the timeline is discretized into a set of time slices, measurements of the system state taken at regularly-spaced intervals. The graph G can now state dependencies across multiple slices, representing dependencies across time. In this case, the joint probability distribution of the system is:

$$P(\chi(0:N))=\prod_{n=0}^{N-1} P(\chi(n+1)|\chi(0:n)) \quad (16)$$

where $\chi_{(n)}$ is the set of variables that represent the system state at time n. Assuming conditional independence of the future with respect to the past given the present state of the system (also known as the Markov assumption):

$$P(\chi(0:N))=\prod_{n=0}^{N-1} P(\chi(n+1)|\chi(n)) \quad (17)$$

If we further assume that the dynamic system is stationary, then any time dependence is the same for all n. In this case, the system can be represented using just the initial state distribution and the transition model $P(\chi(n)|\chi(n-1))$. One such case of a DBN is the Hidden Markov Model. Despite their simplicity, HMM are an extremely useful architecture, with applications in speech recognition systems and analysis of biological sequences.

4. Experiments

4.1. Classification of simulated data using SICAMM

The performance of SICAMM was tested using a simple simulation, similar to the first example provided in a classical reference [19]. Data for this simulation were gener-

ated using a two-class Sequential ICAMM, with two samples per observation. The results obtained from ICAMM and SICAMM will be compared with the classification obtained with a DBN that implemented a HMM structure with mixtures of Gaussians to model the probability density function of data.

The mixing matrices for each class were randomly generated from a uniform distribution in the range [0, 1]. The centroids were selected relatively close, thus $\mathbf{b}_1 = [1\ 1]^T$ and $\mathbf{b}_2 = [1.5\ \ 1.5]^T$. The sources were uniform (for class 1) and Laplacian (for class 2), having zero mean and unit variance. The classes $C_k(n)$ were randomly generated from a HMM structure with a symmetric transition matrix. Then, only one parameter α is required to establish the degree of a sequential dependence, since

$$P(C_1(n)|C_1(n\text{-}1)) = P(C_2(n)|C_2(n\text{-}1)) = \alpha$$

$$P(C_1(n)|C_2(n\text{-}1)) = P(C_2(n)|C_1(n\text{-}1)) = 1\text{-}\alpha$$

A total of 1024 observations were generated for each run of the simulation. The first half of these observations was used to train the classifiers, and the remaining half was used to assess the classification performance of each method. The simulation was repeated 300 times, and the results were averaged.

Figure 1 represents the average classification rate (CR) of each method for α varying from 0.5 (no sequential dependence at all) to 1 (total dependence). The best result is achieved by SICAMM, with the second best result obtained either by ICAMM (for low dependence) or DBN (for high dependence). Notice that the CR of SICAMM and DBN increase for α>0.7, while the CR of ICAMM remains at a constant level. This is an obvious consequence of the sequential dependence of successive classes in the generative model. Notice too that SICAMM and ICAMM achieve very similar results for low α, which means that SICAMM is still able to perform correctly even without sequential dependence.

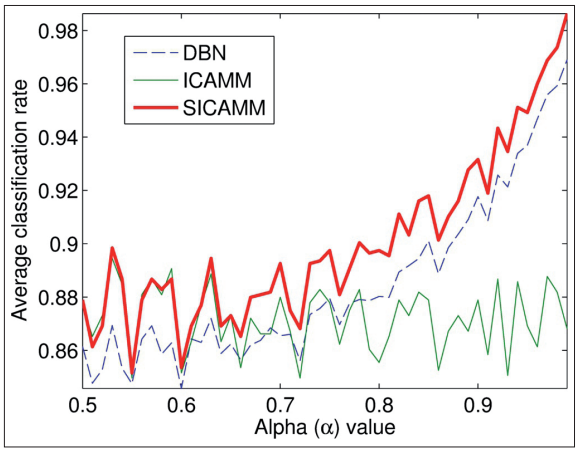

Figure 1. *Average classification rate of the considered methods.*

SICAMM models the statistical dependence between two successive time instants using Hidden Markov Models. PREDICAMM uses known data and ICA Mixture Models to estimate missing data.

4.2. Classification of real data using SICAMM

In this subsection we show a practical application of ICAMM and SICAMM to a real data problem, in the field of computer-assisted sleep staging [44]. Human sleep can be split in four different stages: wake, light sleep, deep sleep, and rapid-eye movement sleep (REM, also known as paradoxical sleep). The sequence of these stages can be used to diagnose different sleep disorders, illnesses (e.g. depression), and some drug addictions. Of particular importance is the detection of very short periods of wakefulness [45], also called arousals, since their rate of appearance can help with the diagnosis of apnea and epilepsy.

The record of the different sleep stages corresponding to a given period is called a hypnogram. Hypnograms are usually obtained in a non-automated manner, by visual inspection of an expert of the polysomnogram (PSG), a set of EEG and other biological records obtained from the sleeping patient. There have been some advances towards the automation of the hypnogram [44], but a totally automatic system remains a challenge.

In this work, ICAMM and SICAMM were applied to perform automatic detection of arousals. Hence, the hypnograms will only show two stages: stage 1, corresponding to any stage of sleep; and stage 2, corresponding to arousals. Once again, the results from ICAMM and SICAMM will be compared against those of a DBN. These automatic classifications will be compared with the non-automatic detection made by an expert.

Two patients with apnea were considered for the experiment, and their records were approximately 8 hours long. These records, and the reference hypnograms, were obtained by an expert using conventional non-automatic procedures.

Each PSG was composed by 24 channels, 11 EEG channels and 13 other channels measuring muscle signals, breathing, and other physiological characteristics. These signals were split in short epochs, typically 1-3 seconds long, and for features were calculated at each epoch. These features where then averaged for 30-second epochs. Thus, a decision about the sleep stage is taken every 30 seconds. The four features extracted from the PSG signals were: amplitude, dominant rhythm, and theta-slow-wave index (TSI) from channel C3-A2; and alpha-slow-wave index (ASI) from EEG channel O2-A1. The dominant rhythm was estimated as the pole frequency of the second-order autoregressive (AR) model; the ASI was the power ratio in the alpha band (8-11 Hz) to the combined power in the delta (0.5-3.5 Hz) and theta (3.5-8 Hz) bands; and the TSI was the ratio of power in the theta band to the combined

power in the delta and alpha bands. These features are commonly used in PSG analysis [44][45].

The first half of the data was used to estimate the parameters of each classifier (ICAMM, SICAMM and BN) in a supervised form, considering the labeling provided by the expert. The ICAMM parameters, $\mathbf{W}_k = \mathbf{A}_k^{-1}$, $\mathbf{s}_k(n)$, $\mathbf{b}_k$, $k=1,\ldots K$ were estimated from the training record in a supervised form using the JADE ICA algorithm [46]. The probabilities of transition between stages were also estimated from the training record, see Table 1. Note that the probabilities of permanence in the same class are clearly above 0.5, justifying the use of sequential dependence in this application.

Subject 1		
	Stage 1	Stage 2
Stage 1	0.98	0.02
Stage 2	0.31	0.69

Subject 2		
	Stage 1	Stage 2
Stage 1	0.96	0.04
Stage 2	0.20	0.80

Table 1. *Estimated transition probabilities, $P(C_k(n)/C_l(n\text{-}1))$, for the studied subjects.*

Subject number	ICAMM		SICAMM		BN	
	CN	NCR	CR	NCR	CR	NCR
1	79,78	69.02	91.01	74.92	89.66	69.68
2	63.66	54.70	79.74	62.04	76.87	51.01

Table 2. Classification rate (CR) and normalized classification rate (NCR) for the two examined subjects.

Table 2 shows the classification rate (CR) and normalized classification rate (NCR) for each one of the considered classifiers. The NCR is the average of the true positive rate and the true negative rate, which helps compensate for the highest proportion of stage-1 epochs. SICAMM achieves the best result, having the highest CR and NCR, while the Bayesian Network is second. Both methods achieve better results than ICAMM, which is to be expected since they consider sequential dependence. The classifications achieved for each subject are shown in Fig. 2 (patient 1) and Fig. 3 (patient 2), for better comparison against the classification estimated by the expert.

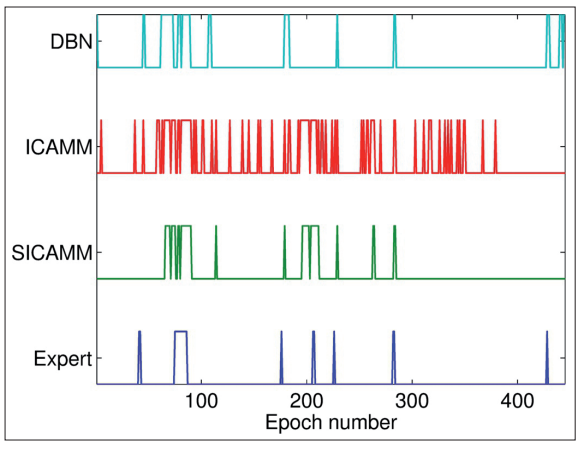

Figure 2. *Arousal detection for subject 1.*

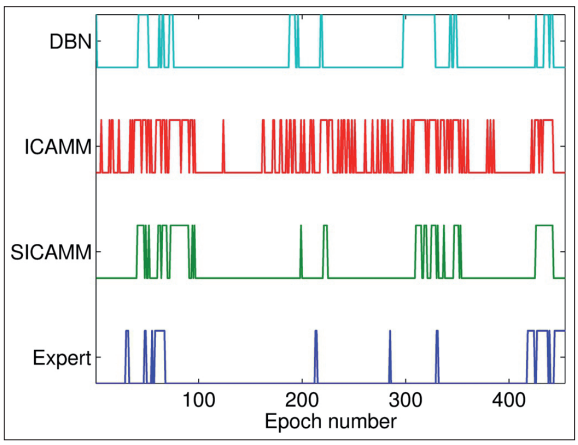

Figure 3. *Arousal detection for subject 2.*

4.3. Prediction using PREDICAMM

In this subsection we present the results of PREDICAMM for the prediction of real data from an EEG experiment. The EEG data was recorded from subjects performing the Sternberg task, a classic test of multi-item short-term memory [47]. Fig. 4 shows an outline of the stages in a Sternberg working memory task. During this task, each participant is shown series (or trials) of one to five symbols (or items), each one taken at random from a fixed set. Each symbol is displayed during 0.2 seconds (item stage), after which the screen is cleared for 1 second (pause stage), and then the following symbol appears on screen. The length of the series varies at random from trial to trial. After the last symbol, there is a further 1-second delay (retention stage), a warning signal, and then a test symbol. The subject is then required to decide whether or not the test symbol is one of the symbols shown in this series (probe stage). Positive and negative responses are required with equal frequency. Finally, after the subject has decided, there is a pause of 0.5 s until the next trial. For this particular experiment, each task comprised 30 trials. There were a total of three subjects, with 2 to 5 experiments per subject. The subjects showed a success rate of 98 %, with an average response time of 1.17 s.

EEG signals were recorded from electrodes using water-based gel coupling and placed according to the 10-10 system. Sixty-four channels were recorded at 512 Hz, using Cz as reference. The signals were recorded using a Biosemi device with active electrodes. All channels were band-pass filtered between 1 and 70 Hz, with an additional narrow notch filter at 50 Hz. To estimate the ICA mixture model in the EEG data, we used the on-line ICAMM algorithm shown in [35].

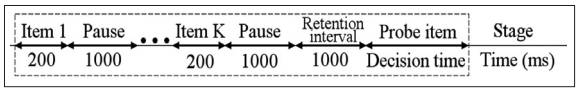

Figure 4. *Description of the Sternberg task.*

In order to test the predictive capabilities of the method, a case was considered where the values at one or more of the electrodes were unusable and had to be predicted. This would be the case, e.g. if some of the electrodes had

been disconnected, or for artifact-removal purposes. The performance of PREDICAMM was compared against sphere splines, a classical method for EEG prediction [48]. This prediction was performed by removing one or more of the EEG channels (chosen at random) and estimating the missing data. This process was repeated 1000 times, with the end result being the average prediction error.

Three figures of merit or indices were used to evaluate the quality of the prediction: Signal-to-Interference Ratio (SIR), the Kullback-Leibler divergence (KLD), and the Mean Structural Similarity (MSSIM). SIR measures the prediction error and is, essentially, the inverse of the mean squared error: the higher the SIR, the better the prediction. KLD, on the other hand, measures the distance between the probability densities of predicted and true data. Finally, normalized MSSIM is an index, commonly used in image processing, that measures the structural similarity between predicted and true data [49]. The closer the MSSIM value to 1, the higher the similarity.

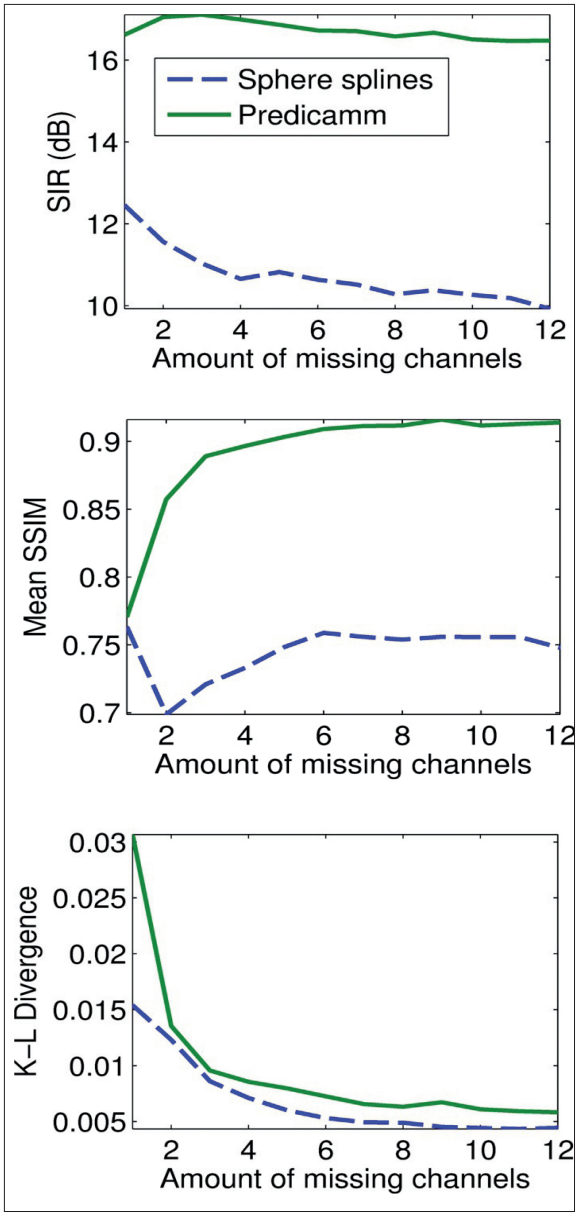

Figure 5. *Average error indicators for the considered predictors.*

Of particular importance is the detection of very short periods of wakefulness, also called arousals, since their rate of appearance can help with the diagnosis of apnea and epilepsy.

Fig. 5 shows the average prediction results for an increasing number of missing channels for sphere splines and for PREDICAMM Sphere splines achieve a better KLD, which means that the distribution of the predicted data is more similar to that of true data. This was to be expected, since sphere splines have been used to model surface EEG data. On the other hand, PREDICAMM obtains much better SIR and MSSIM. PREDICAMM achieves this improved result by modeling the local behavior of the EEG signals through the ICAMM.

It is notable that the prediction of both methods seems to increase for some indicators when the number of missing channels is increased. This effect is found because higher-amplitude channels (e.g. frontal channels) are easier to interpolate correctly than lower-amplitude ones. Since the missing channels were chosen at random, a larger number of missing channels implies a higher chance of selecting at least one frontal channel, thus increasing average performance.

To further showcase the performance of PREDICAMM, Fig. 6 shows the prediction for a given time instant from a case with 32 missing channels. It can be seen that PREDICAMM achieves a much lower prediction error overall, and the prediction is more similar to true data.

5. Conclusions

Two applications of ICAMM to the processing of EEG signals were proposed, one for prediction (PREDICAMM) and one for classification (SICAMM). Estimates of the ICAMM parameters are required for both and have to be estimated from training data.

The first application, SICAMM, is an extension of the basic ICA Mixture Model that includes information about the sequential dependence of the classes. The performance of SICAMM has been compared against ICAMM and a Bayesian Network, both with simulations and with real EEG data. Results show the potential of SICAMM for the classification of EEG signals, with SICAMM outperforming the results of the Bayesian Network.

The second application, PREDICAMM, makes use of the ICAMM parameters to estimate missing data. PREDICAMM was used to predict missing data from real EEG data, and compared against the prediction from a classical predictor for EEG signals, sphere splines. The results of PREDICAMM clearly outperformed the classic method, since ICA Mixture Models allow for greater flexibility, modeling local nonlinearities while keeping the general structure of the data, and thus improving the result of the prediction.

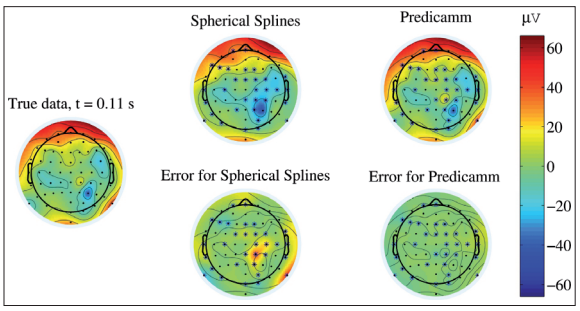

Figure 6. *Scalp maps of the prediction and prediction error for a case with 32 missing channels. Missing channels are indicated by stars (*), while known channels are marked by dots (·).*

Taking into account the success of ICA application to EEG signal processing, the flexibility of ICAMM as a non-linear extension of ICA suggest it for future EEG applications. Further work is required to develop algorithms which can simultaneously estimate all the model parameters in an unsupervised framework.

Acknowledgment

This work has been supported by Universitat Politècnica de Valencia under grant 20130072, Generalitat Valenciana under grants PROMETEO/2010/040 and ISIC/2012/006; and Spanish Administration and European Union FEDER Programme under grant TEC2011-23403 01/01/2012. The PSG signals and annotated hypnograms were provided by the Electroencephalography Department of Hospital Universitario La Fe, Valencia, Spain.

References

[1] E. Niedermeyer and F.L. da Silva, Electroencephalography: Basic Principles, Clinical Applications, and Related Fields. Lippincot Williams & Wilkins, USA, 2004.

[2] P. Common, and C. Jutten, Handbook of Blind Source Separation: Independent Component Analysis and Applications. Academic Press, USA, 2010.

[3] A. Salazar, On Statistical Pattern Recognition in Independent Component Analysis Mixture Modelling, Springer Theses in Electrical Engineering Series, Springer-Verlag, Berlin, Heidelberg, 2013.

[4] A. Salazar, and L. Vergara, Perspectives on pattern recognition from ICA mixture modeling, In Perspectives on Pattern Recognition, Ed. Fournier, M.D., Nova Science Publishers, Inc, New York, pp. 203-223, 2012.

[5] A. Salazar and L. Vergara, Knowledge discovery from E-Learning activities. In Advances in E-Learning: Experiences and Methodologies, Ed. F. Garcia-Peñalvo, IGI-Global Information Science Reference, pp. 173-198, 2008.

[6] V. Zarzoso, O. Meste, P. Comon, D. G. Latcu and N. Saoudi, Noninvasive cardiac signal analysis using data decomposition techniques, in: F. Cazals and P. Kornprobst (Eds.), Modeling in Computational Biology and Biomedicine: A Multidisciplinary Endeavor, Berlin, Heidelberg: Springer Verlag, chapter 3 pp. 83-116, 2013.

[7] V. Zarzoso and A. K. Nandi, "Noninvasive fetal electrocardiogram extraction: blind separation versus adaptive noise cancellation," IEEE Transactions on Biomedical Engineering, vol. 48, no. 1, January, pp. 12-18, 2001.

[8] C.J. James, and C.W. Hesse, "Independent component analysis for biomedical signals," Physiological Measurement, vol. 26, pp. R15-R39, 2005.

[9] A. Cichocki, Generalized Component analysis and blind source separation methods for analyzing mulitchannel brain signals, statistical and process models for cognitive neuroscience and aging, Lawrence Erlbaum Associates, pp. 201-272, 2007.

[10] R. Llinares, J. Igual, A. Salazar, and A. Camacho, "Semi-blind source extraction of atrial activity by combining statistical and spectral features," Digital Signal Processing: A Review Journal, vol. 21 no.2, pp. 391-403, 2011.

[11] R. Llinares, J. Igual, A. Salazar, J. Miro-Borras, and A. Serrano, "Atrial activity extraction based on statistical and spectral features," Lecture Notes in Computer Science, vol.5441, pp. 451-458, 2009.

[12] R. Llinares, J. Igual, A. Salazar, and L. Vergara, "Constrained temporal extraction of the atrial rhythm in Atrial Fibrillation episodes," in Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society, 2008, pp. 1159-1162.

[13] C. Klein and B. Feige, "An independent component analysis (ICA) approach to the study of developmental differences in the saccadic contingent negative variation," Biological Psychology, vol. 70, pp. 105-114, 2005.

[14] S. Makeig, M. Westerfield, T.P. Jung, J. Covington, J. Townsend, T.J.Sejnowski and E. Courchesne, "Functionally independent components of the late positive event-related potential during visual spatial attention," Journal of Neuroscience, vol. 19, no. 7, pp. 2665-2680, 1999.

[15] M. Wibral, G. Turi, D.E.J. Linden, J. Kaiser, and C. Bledowski, "Decomposition of working memory-related scalp ERPs: Crossvalidation of fMRI-constrained source analysis and ICA," Internt J. of Psychol, vol. 67, pp. 200-211, 2008.

[16] G. Safont, A. Salazar, A. Soriano, and L. Vergara, "Combination of multiple detectors for EEG based biometric identification/authentication," in Proceedings International Carnahan Conference on Security Technology, 2012, pp. 230-236.

[17] A. Soriano, L. Vergara, G. Safont, and A. Salazar, "On comparing hard and soft fusion of dependent detectors," in Proceedings 22nd IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2012; Santander; Spain; 23 September, 2012.

[18] N.P. Castellanos and V.A. Makarov, "Recovering EEG brain signals: Artifact suppression with wavelet enhanced independent component analysis," Journal of Neuroscience Methods, vol. 158, pp. 300-312, 2006.

[19] T.W. Lee, M.S. Lewicki, and T.J. Sejnowski, "ICA mixture models for unsupervised classification of non-

gaussian classes and automatic context switching in blind signal separation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 10, pp. 1078-1089, 2000.
[20] A. Salazar, L. Vergara, A. Serrano, and J. Igual, "A general procedure for learning mixtures of independent component analyzers," Pattern Recognition, vol. 43, no. 1, pp. 69-85, 2010.
[21] A. Salazar, and L. Vergara, "ICA mixtures applied to ultrasonic nondestructive classification of archaeological ceramics", Eurasip Journal on Advances in Signal Processing, vol. 2010, article ID 125201, 11 pages, 2010.
[22] A. Salazar, A. Rodriguez, G. Safont, and L. Vergara, "Prospective of the application of ultrasounds in archaeology," Materials Science and Engineering, vol. 42, no.1, pp. 1-4, 2012.
[23] A. Salazar. L. Vergara, and R. Llinares, "Learning material defect patterns by separating mixtures of independent component analyzers from NDT sonic signals," Mechanical Systems and Signal Processing, vol. 24, no.6, pp. 1870-1886, 2010.
[24] A. Salazar, A. Serrano, R. Llinares, L. Vergara, and J. Igual, "ICA mixture modeling for the classification of materials in impact-echo testing," Lecture Notes in Computer Science , vol. 5441, pp. 702-709, 2009.
[25] A. Salazar, J. Igual, L. Vergara, and A. Serrano, "Learning hierarchies from ICA mixtures," in Proceedings of IEEE International Conference on Neural Networks , 2007, pp. 2271-2276.
[26] A. Salazar, L. Vergara, and R. Miralles, "On including sequential dependence in ICA mixture models," Signal Processing, vol. 90, pp. 2314-2318, 2010.
[27] P. Dayan and L.F. Abbot, Theoretical neuroscience: computational and mathematical modeling of neural systems. The MIT Press, USA, 2001.
[28] G. Safont, A. Salazar, L. Vergara, A. Gonzalez, and A. Vidal, "Mixtures of independent component analyzers for EEG prediction," Communications in Computer and Information Science, vol. 338, pp. 328-335, 2012.
[29] G. Safont, A. Salazar, A. Rodriguez, and L. Vergara, "New applications of sequential ICA mixtures models compared with dynamic Bayesian networks for EEG signal processing," in Proceedings of IEEE 5th International Conference on Computational Intelligence, Communication Systems and Networks, CICSyN2013, pp. 397-402, Madrid, 2013.
[30] A. Rodríguez, A. Salazar, and L. Vergara, "Analysis of split-spectrum algorithms in an automatic detection framework," Signal Processing, vol. no. 9, pp. 2293-2307, 2012.
[31] A. Rodríguez, A. Salazar, and L. Vergara, "Defect characterization in steel alloys using the modified split-spectrum algorithm," In Proceedings of AIP, vol. 1433, 2012, pp. 483-486.
[32] G. Safont, A. Salazar, L. Vergara, R. Llinares, and J. Igual, "Wiener systems for reconstruction of missing seismic traces," in Proceedings of the International Joint Conference on Neural Networks, 2011, pp. 1049-1053.
[33] G. Safont, A. Salazar, and L. Vergara, "Nonlinear prediction based on independent component analysis mixture modelling," Lecture Notes in Computer Science, vol. 6692 LNCS (PART 2), pp. 508-515, 2011.
[34] J.M. Gorriz, C.G. Puntonet, G. Salmeron, and E.W. Lang, "Time series prediction using ICA algorithms," in Proceedings of 2nd IEEE International Workshop on Intelligent Data Acquisition and Advanced Computer Systems: Technology and Applications, 2003, pp. 226-230.
[35] C.T. Lin, W.C. Cheng, and S.F. Liang, "An On-line ICA-Mixture-Model-Based Self-Constructing Fuzzy Neural Network," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 52, no. 1, pp. 207-221, January 2005.
[36] C.A. Shah, P.K. Varshney, and M.K. Arora, "ICA mixture model algorithm for unsupervised classification of remote sensing imagery," International Journal of Remote Sensing, vol. 28, no. 8, pp. 1711–1731, 2007.
[37] T.W. Lee, M.S. Lewicki, "Unsupervised image classification, segmentation, and enhancement using ICA mixture models," IEEE Transactions on Image Processing, vol. 11, no. 3, pp. 270–279, 2002.
[38] O. Cappe, E. Moulines, and T. Ryden, Inference in Hidden Markov Models. Springer, New York, 2005.
[39] D. Koller and N. Friedman, Probabilistic Graphical Models: Principles and Computation. The MIT Press, USA, 2009.
[40] B.J. Wiliams and B. Cole, "Mining monitored data for decision-making with a Bayesian network model," Ecological modelling, vol. 249, pp.26-36, 2013.
[41] M. Grzegorczyk, D. Husmeier, and J. Roahnenführer, "Modelling non-stationary dynamic gene regulatory processes with the BGM model," Computational Statistics, vol. 26 no. 2, pp.199-218, 2011.
[42] L. Hausfeld, F. De Martino, M. Bonte, and E. Formisano, "Pattern analysis of EEG responses to speech and voice: Influence of feature grouping," Neuroimage, vol. 59, no.4, pp. 3641-3651, 2012.
[43] A. Chan, E. Halgren, K. Marinkovic, and S. Cash, "Decoding word and category-specific spatiotemporal representations from MEG and EEG," Neuroimage, vol. 54, no. 4, pp. 3028-3039, 2011.
[44] R. Agarwal, and J. Gotman, "Computer-assisted sleep staging," IEEE Transactions on Biomedical Engineering, vol. 48, no. 12, pp. 1412–1423, 2001.
[45] M. Jobert, H. Shulz, P. Jähnig, C. Tismer, F. Bes, and H. Escola, "A computerized method for detecting episodes of wakefulness during sleep based on the Alpha slow-wave index (ASI)," Sleep, vol. 17, no.1, pp. 37–46, 1994.
[46] J.F. Cardoso, and A. Souloumiac, "Blind beamforming for non-Gaussian signals," IEE Proceedings-F, vol. 140, no. 6,362–370, 1993.
[47] S. Sternberg, "High-speed scanning in human memory," Science, vol. 153, no. 3736, pp. 652-654, 1966.
[48] F. Perrin, J. Pernier, D. Bertrand, and J.F. Echallier, "Spherical splines for scalp potential and current density matching," Electroencephalography and Clinical Neurophysiology, vol. 72, pp. 184-187, 1989.
[49] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13 no. 4, pp. 600-612, 2004.

Biographies

MSc. Gonzalo Safont received his B.Sc. degree and his M.Sc. degree in Telecommunications Engineering from the Universidad Politécnica de Valencia (UPV) in 2008 and 2011, respectively. He is a Ph.D. student at the Institute of Telecommunication and Multimedia Applications of UPV. He is currently researching on advanced methods for prediction, and dynamic modeling based on mixture of independent component analyzers and decision fusion techniques. He has worked in different applied problems including non-destructive testing, and biomedical diagnosis.

Dr. Addisson Salazar received the B.Sc. and M.Sc. degrees in Information and Systems Engineering from Universidad Industrial de Santander, the D.E.A. degree in Telecommunications from Universidad Politécnica de Valencia (UPV) in 2003, and the Dr. in Telecommunications degree from UPV in 2011. He is a senior research official in the Institute of Telecommunications and Multimedia Applications at UPV since 2007. His research interests include statistical signal processing, machine learning, and pattern recognition with emphasis on methods for signal classification based on decision fusion, time-frequency techniques, and mixtures of independent component analyzers. The application of his research has been focused on data mining, nondestructive testing and biomedical problems.

Dr. Alberto Rodríguez received the Telecommunication Engineer degree from the Universidad de Vigo in 1998 and the Dr. Engineer in Telecommunications degree from the Universidad Politecnica de Valencia in 2011. He is an Associate Professor in the Communications Engineering Department in the Universidad Miguel Hernandez de Elche. His research focuses on statistical signal processing, time-frequency analysis techniques and signal detection and identification. The applications of his research are focused in ultrasound signal processing, biomedical signal processing, brain-to-computer interfaces and LADAR image processing.

Prof. Luis Vergara received the Telecommunication Engineer and the Dr. Engineer of Telecommunication degrees from the Universidad Politécnica de Madrid (UPM) in 1980 and 1983, respectively. Since 1992, he is a Professor at the Communication Department (Universidad Politécnica de Valencia, Spain). His research concentrates in the statistical signal processing area, where he has worked in different theoretical an applied problems, many of them under contract with the industry. His theoretical aspects of interest are signal detection, classification, decision fusion, independent component analysis and spectral analysis. Currently he is involved in ultrasound signal processing for non-destructive evaluation, in infrared signal processing for fire detection, automatic credit card fraud detection, and in cognitive audio for surveillance applications. He has published more than 150 papers including journals and conference contributions.