Multichannel Dynamic Modeling of Non-Gaussian Mixtures

Gonzalo Safont[(1)], Addisson Salazar[(1)], Luis Vergara[(1)], Enriqueta Gómez[(2)], Vicente Villanueva[(2)]

[(1)] Institute of Telecommunications and Multimedia Applications, Universitat Politècnica de València, Spain

[(2)] Hospital Universitari i Politècnic La Fe, Valencia, Spain

Corresponding author: Gonzalo Safont, gonsaar@upvnet.upv.es

***Abstract*— This paper presents a novel method that combines coupled hidden Markov models (HMM) and non-Gaussian mixture models based on independent component analyzer mixture models (ICAMM). The proposed method models the joint behavior of a number of synchronized sequential independent component analyzer mixture models (SICAMM), thus we have named it generalized SICAMM (G-SICAMM). The generalization allows for flexible estimation of complex data densities, subspace classification, blind source separation, and accurate modeling of both local and global dynamic interactions. In this work, the structured result obtained by G-SICAMM was used in two ways: classification and interpretation. Classification performance was tested on an extensive number of simulations and a set of real electroencephalograms (EEG) from epileptic patients performing neuropsychological tests. G-SICAMM outperformed the following competitive methods: Gaussian mixture models, HMM, Coupled HMM, ICAMM, SICAMM, and a long short-term memory (LSTM) recurrent neural network. As for interpretation, the structured result returned by G-SICAMM on EEGs was mapped back onto the scalp, providing a set of brain activations. These activations were consistent with the physiological areas activated during the tests, thus proving the ability of the method to deal with different kind of data densities and changing non-stationary and non-linear brain dynamics.**

***Index Terms*—dynamic modeling, non-Gaussian mixtures, ICA, HMM, EEG**

I. Introduction

A. Background

In a simple pattern recognition statement, a given unknown state must be determined from a vector of observations. This evolves towards more complex settings where a sequence of states must be estimated considering not only the observed data but also the state dynamics (dependence among states at different times). Hidden Markov models (HMM, [1]) is by far the most popular approach to estimate hidden sequential states from observed data. This is due to its reasonable simplicity in conjunction with good capability to capture local state dynamics (dependences between the states corresponding to consecutive time instant).

On the other hand, observed data and states are related by the conditional probability density function (pdf). The usual assumption is that the conditional pdf is multivariate Gaussian, because the model parameters can be efficiently estimated. This implies that the unconditional pdf of the observed data fits a Gaussian mixture model where every component of the mixture is the state conditional pdf. But Gaussianity relates to linearity which is not a reasonable assumption in many real life problems. Some examples where non-linearity in the probability has been considered are: action recognition via sparse Gaussian processes [2]; modeling growth dynamics using unscented Kalman filters [3]; an extended Kalman filter augmented with local searches [4]; and modeling the data using a two-step method with fuzzy clustering and Gaussian mixture models (GMM) [5]. Some particular non-Gaussian conditional probabilities have been proposed in HMMs in applications such as handwritten word recognition [6] and biological sequence analysis [7].

A general extension of GMM to non-Gaussian mixtures is based on the concept of independent component analyzers (ICA). Briefly, ICA is a blind source separation technique that separates a set of observations into a group of non-Gaussian and statistically independent sources [8, 9]. We can model every conditional probability as an ICA and then the unconditional probability of the data will be an ICA mixture model (ICAMM) [10]. In ICAMM, the source model is assumed to have several hidden states, and the data from each state are modeled by a different ICA. A general implementation of ICAMM included non-parametric source estimation and semi-supervised learning [11-13]. In [14, 15], ICAMM was proposed to model the conditional pdfs in a HMM, the method was called sequential ICA mixture models (SICAMM). ICA mixture emission distribution in each hidden state was assumed.

B. Motivation and paper organization

The main contribution of the paper is the generalization of SICAMM to multiple coupled HMM, that we will call generalized SICAMM (G-SICAMM). Use of multiple coupled HMM is a more efficient way to deal with the observed data when they come from different channels. This may happen when we want to combine data from different modalities and/or data captured from different sensors. It is true that we could combine the observed data vectors of the different channels in just one data vector. However, notice that the number of model parameters to be estimated increases very significantly. Thus, for example, let us consider 2 coupled HMM, where the dimension of the observation vectors is M. In SICAMM we have to estimate a mixing matrix of dimension $2M \times 2M$, a total of $4M^2$ parameters, while in G-SICAMM we have to estimate two mixing matrices of dimension $M \times M$, a total of $2M^2$ parameters.

There are precedents of using coupled HMM (CHMM). CHMM is a state of the art method for dynamic modeling that has been implemented in Bayesian networks using GMM [16]. It has been used in several pattern recognition applications, such as modeling intra-modal dependences in multimodal data for video-realistic speech animation [17] and sign language recognition [18]. In contrast with GMM-based methods, the non-Gaussianity of the sources extracted by G-SICAMM allows for source identification and facilitates their interpretation and association with meaningful variables of real applications. G-SICAMM is suitable for a myriad of scenarios where local interaction dynamics have to be modeled while preserving global relationships between the variables. This condition is inherent to several problems where topological and/or functional setups determine the requirement of partitioning in several Markov chains. Depending on the application, the partition could be related with physical components, sensor spatial locations, or data modalities. Therefore, G-SICAMM allows for decomposition of a particular phenomenon for subspace analyses, bearing in mind physical model understanding and/or data mining exploitation. With an adequate preprocessing of the observations, G-SICAMM can be used in any classification or pattern recognition problem. It would be potentially useful when the objective is to analyze long term non-stationary processes with nonlinearities and temporal dependence.

All in all, from a motivation perspective, it can be concluded that realizing the inner structure of the observed data emanating from the multichannel setting makes G-SICAMM a more efficient method to deal with a limited amount of data for both training and testing in comparison with one channel methods like SICAMM or HMM. It is also more versatile than other coupled methods like CHMM, as non-Gaussian modelling is possible. Finally, the inherent ICA structure of G-SICAMM may lead to a better interpretation of the results by analyzing the estimated independent sources that are originating the data.

We have considered the following competitive methods for comparison: GMM, HMM (GMM for the observations pdf), CHMM (GMM for the observations pdf), ICAMM, SICAMM, and long short-term memory (LSTM) recurrent neural networks [19, 20]. An extensive number of simulations to evaluate the dynamic modeling performance were made. The results demonstrated the capabilities of G-SICAMM to exploit the sequential dependence of successive states in the same Markov chain and the temporal dependence of the global state among multiple Markov chain interactions. A real application consisting of classification of stages of neuropsychological tests using electroencephalographic (EEG) signals was also considered. Two tests with different kinds of stimuli were implemented: the visual memory Barcelona Neuropsychological Test (TB) [21] and TAVEC [22], an auditory working memory test. Those tests were being performed by epileptic patients as part of their clinical diagnosis. The classification results of G-SICAMM outperformed those of all the other competitive methods in terms of the balanced error rate (BER) and the recall measured by Cohen's kappa coefficient [23]. Furthermore, patterns of physiological significance were obtained in the parameters of G-SICAMM.

The rest of the paper is composed by the following parts. Section 2 reviews the background of SICAMM; Section 3 includes the development of the proposed G-SICAMM method; Section 4 contains the explanations of the simulations; Section 5 is devoted to the real application; and finally, Section 6 includes discussion, conclusion, and future lines of research derived from this work.

II. Sequential ICA Mixtures

We will use matrix notation for the variables defined in this work. For clarity of notation, we will denote random variables and their realizations with the same symbols, and the difference between both can be deduced from context.

A. Independent Component Analyzers

Let us assume that we have a set of *N* observations, $\mathbf{x}(n)$, $n = 1...N$. For simplicity, we will assume that these observations are centered. Independent component analyzers ([8]) search simultaneously for a mixing matrix $\mathbf{A}$ and a set of independent sources $\mathbf{s}(n)$ such that

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

$\mathbf{A}$ is a matrix of size $[M \times R]$, where R is the number of extracted sources at each time instant and M is the number of variables of the observations. For simplicity, we will assume that $R=M$ and that $\mathbf{A}$ can be inverted to find the demixing matrix, $\mathbf{W}=\mathbf{A}^{-1}$. Thus, the sources can be estimated from the observations as $\mathbf{s}(n)=\mathbf{W}\cdot\mathbf{x}(n)$, and individual sources can be estimated as $s_m(n)=\mathbf{w}_m{}^T\cdot\mathbf{x}(n)$, where $\mathbf{w}_m{}^T$ is the *m*th row of $\mathbf{W}$, $1\leq m\leq M$. Due to the independence consideration of ICA, the multivariate probability density function of the observations can be obtained as a product of one-dimensional pdfs:

$$p(\mathbf{x}(n))=|\det\mathbf{W}|\cdot\prod_{m=1}^{M}p_{s_m}(s_m(n)) \tag{2}$$

where $p_{s_m}(\ )$ is the marginal density of the *m*th source.

The use of ICA mixture models (ICAMM) was first proposed in [10], which considered a source model switching between Laplacian and bimodal densities. ICAMM considers a mixture of *K* separate ICA models, each with its own mixing matrix $\mathbf{A}_k$, sources $\mathbf{s}_k(n)$, and center, $\mathbf{b}_k$, $k = 1\ldots K$. It is a switching model, so that, if the nth observation belongs to state k (i.e., $c(n)=k$), the data can be expressed as

$$\mathbf{x}(n)=\mathbf{A}_k\cdot\mathbf{s}_k(n)+\mathbf{b}_k \tag{3}$$

This model increases the flexibility of ICA, because sources are independent from other sources in the same state, but they can have any dependence with sources from different states. Furthermore, each state can be centered around different values. In essence, $\mathbf{A}_k$ and $\mathbf{s}_k(n)$ determine the shape of the "cluster" of points during state *k*, and $\mathbf{b}_k$ determines its center. Given the switching model, the pdf of each observation is estimated as $p(\mathbf{x}(n))=\sum_{k=1}^{K}p(\mathbf{x}(n)\,|\,c(n)=k)\cdot p(k)$, with $p(k)$ being the prior probability of state k and

$$\begin{aligned}p(\mathbf{x}(n)\,|\,c(n)=k)&=|\det\mathbf{W}_k|\cdot p_{\mathbf{s}_k}(\mathbf{s}_k(n))=\\&=|\det\mathbf{W}_k|\cdot\prod_{m=1}^{M}p_{s_{k,m}}(s_{k,m}(n))\end{aligned} \tag{4}$$

where $s_{k,m}(n)=\mathbf{w}_{k,m}{}^T(\mathbf{x}(n)-\mathbf{b}_k)$ is the estimate of the *m*th source at time *n*, given the model was in state *k*.

B. Sequential ICA mixture model

In the case of ICAMM, the likelihood of the data is usually obtained assuming there is no time dependence in the observations nor between states. This assumption simplifies the calculation of the probabilities involved in the process, since in that case the likelihood becomes $p\left(\mathbf{x}(1)...\mathbf{x}(N)\right)=\prod_{n=1}^{N}p\left(\mathbf{x}(n)\right)$. However, there are many practical cases where the observations do not behave in a totally independent manner and instead show some degree of time dependence in the feature observation record.

This dependence is considered in sequential ICAMM (SICAMM) [14, 15]. SICAMM is a hidden Markov model (HMM) whose state emissions are modeled as a non-Gaussian mixture using ICA. Therefore, the probability of emitting any given observation $\mathbf{x}(n)$, while the model is at state k, is the same $p\left(\mathbf{x}(n)\,|\,c(n)=k\right)$ shown in (4).

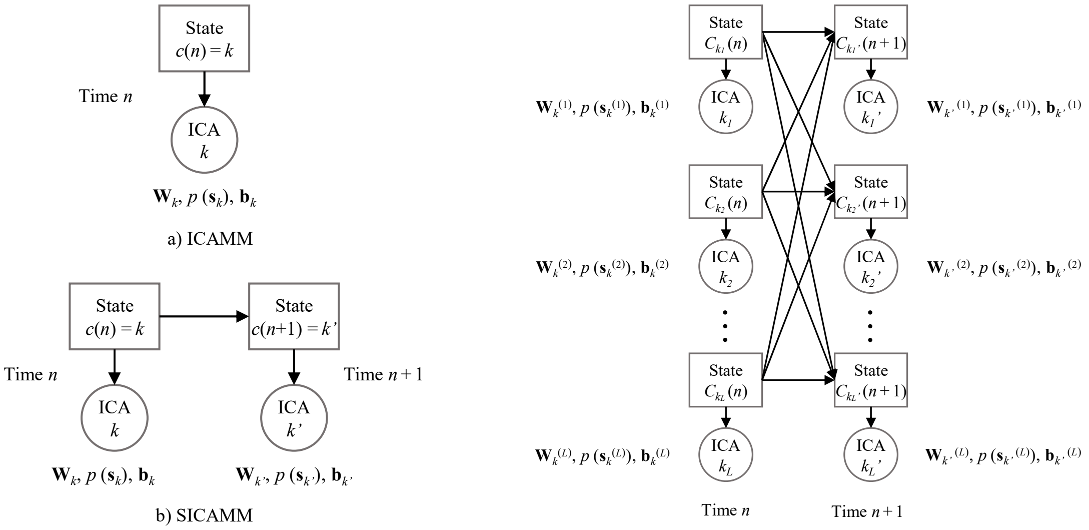
Since each state of the HMM considers a different ICA model, this is, basically, a switching model between different ICAs that incorporate sequential dependence. Thus, SICAMM integrates HMM and ICAMM into a single model. Given the HMM and the emission probabilities in (4), the likelihood of the data can be expressed iteratively, e.g., using classical methods for HMM such as forward-backward (Baum-Welch) and Viterbi [15].

III. Generalized Sequential ICAMM (G-SICAMM)

In this paper, we propose a general framework to characterize the joint behavior of several synchronized SICAMM models, which we have called generalized sequential ICAMM (G-SICAMM). The degrees of freedom of G-SICAMM allow it to accurately model complex local non-Gaussian probability densities and consider time dependencies, without losing global modeling capabilities. It is known that ICA can produce not only a valid statistical model of the data, but also sources with physical or physiological meaning. Examples of this are the extraction of physiologically significant patterns for EEG data [24], the extraction of atrial rhythms during heart fibrillation [25], the removal of physiological artifacts from the EEG signal [26], and the similarities between ICA and image processing in the visual cortex [27]. This capability is inherited by the proposed method. Unlike models that are purely statistical in nature, G-SICAMM can obtain a representation of the data that reflects both their pdf and their underlying generating model. Therefore, the parameters of G-SICAMM could be related with the analyzed physical phenomenon.

G-SICAMM assumes that the observations can be divided into several groups or "chains" of data that behave more or less independently, but whose hidden states are related. There are several possible reasons for this separation, such as: i) the data come from several related, but independently measured, sources (e.g., two different biosignals, such as EEG and ECG [28]); ii) we are interested in isolating the contribution of different input variables. In this work, we will assume the latter, and keep the former for a future work.

In the same way as SICAMM, which integrates HMM and ICAMM, G-SICAMM is a combination of coupled hidden Markov models (CHMM, [16]) and ICA mixture models. G-SICAMM is a CHMM whose emissions are modeled as non-Gaussian mixtures using ICA. Since the emissions of each state of the CHMM are modeled using a different ICA model, G-SICAMM is essentially a synchronized switching model between different ICAs across different chains. Fig. 1.c illustrates the dependencies considered in G-SICAMM. ICAMM has no temporal dependencies (see Fig. 1.a), and SICAMM only has temporal dependence within the same model (see Fig. 1.b). In G-SICAMM (see Fig. 1.c), the state of each model at time n depends on the state of all models at the previous time instant, n - 1. We have considered a fully-coupled model in order to enable asymmetrical and complex dependencies to be modeled. However, other architectures similar to those of factorial HMM [29] or hierarchical HMM [30] could be implemented. Furthermore, we extend the Markov assumption from each individual chain to the joint sets.


State
$c(n)=k$
Time n
ICA
k
$\mathbf{W}_k, p(\mathbf{s}_k), \mathbf{b}_k$
State
$c(n)=k$
State
$c(n+1)=k'$
Time n
Time $n+1$
ICA
k
ICA
k'
$\mathbf{W}_k, p(\mathbf{s}_k), \mathbf{b}_k$
$\mathbf{W}_{k'}, p(\mathbf{s}_{k'}), \mathbf{b}_{k'}$
State
$C_{k_1}(n)$
State
$C_{k_1'}(n+1)$
ICA
k_1
ICA
k_1'
$\mathbf{W}_k^{(1)}, p(\mathbf{s}_k^{(1)}), \mathbf{b}_k^{(1)}$
$\mathbf{W}_{k'}^{(1)}, p(\mathbf{s}_{k'}^{(1)}), \mathbf{b}_{k'}^{(1)}$
State
$C_{k_2}(n)$
State
$C_{k_2'}(n+1)$
ICA
k_2
ICA
k_2'
$\mathbf{W}_k^{(2)}, p(\mathbf{s}_k^{(2)}), \mathbf{b}_k^{(2)}$
$\mathbf{W}_{k'}^{(2)}, p(\mathbf{s}_{k'}^{(2)}), \mathbf{b}_{k'}^{(2)}$
State
$C_{k_L}(n)$
State
$C_{k_L'}(n+1)$
ICA
k_L
ICA
k_L'
$\mathbf{W}_k^{(L)}, p(\mathbf{s}_k^{(L)}), \mathbf{b}_k^{(L)}$
$\mathbf{W}_{k'}^{(L)}, p(\mathbf{s}_{k'}^{(L)}), \mathbf{b}_{k'}^{(L)}$
Time n
Time $n+1$

a) ICAMM

b) SICAMM

c) G-SICAMM

Fig. 1. Comparison of the different dependencies considered in: a) ICA mixture models; b) sequential ICA mixture models; c) generalized sequential ICA mixture models. Square blocks represent the hidden states, round blocks represent the parameters of the emissions, and arrows show dependence between blocks. The *(l)* superindex in c) denote the parameters of the *l*th chain: their definition is otherwise identical to those in Section II.

Before tackling the model itself, we will define several notations which will ease the theoretical development of G-SICAMM. Let us assume that the data are split into L chains. We will denote the parameters of the lth chain with an $^{(l)}$ superscript, $l=1...L$. The observation from the lth chain at time n is denoted by $\mathbf{x}^{(l)}(n)$, and it is a random vector of size $M^{(l)}$. The set of observations from all chains at time n is denoted by random column vector $\mathbf{x}(n)=\left[\mathbf{x}^{(1)}(n)^T, \mathbf{x}^{(2)}(n)^T, ..., \mathbf{x}^{(L)}(n)^T\right]^T$. Each chain is assumed to have been modeled by an ICAMM with $K^{(l)}$ states whose parameters are: $\mathbf{W}_k^{(l)}$, $\mathbf{s}_k^{(l)}(n)$, $\mathbf{b}_k^{(l)}$, $k=1...K^{(l)}$. Finally, we define the random state vector $\mathbf{c}(n)=\left[c^{(1)}(n), c^{(2)}(n)..., c^{(L)}(n)\right]^T$, which contains the state of each one of the L chains at time n, with $1 \le c^{(l)}(n) \le K^{(l)}$. Realizations of the state vector will be denoted by $\mathbf{k}=\left[k^{(1)}, k^{(2)}, ..., k^{(L)}\right]^T$, and the transition probability between the combinations of states $\mathbf{k}$ and $\mathbf{k}'$, $p\left(\mathbf{c}(n)=\mathbf{k}' \mid \mathbf{c}(n-1)=\mathbf{k}\right)$, is denoted by $\pi_{\mathbf{kk}'}$. The prior probability of the combination of states $\mathbf{k}$ is denoted by $p(\mathbf{k})$. Finally, in order to simplify the model, the conditional independence of the data is extended from single chains to the whole G-SICAMM chain structure, i.e., $p\left(\mathbf{x}(n) \mid \mathbf{c}(n)=\mathbf{k}\right)=\prod_{l=1}^{L} p\left(\mathbf{x}^{(l)}(n) \mid c^{(l)}(n)=k^{(l)}\right)$.

In practice, the G-SICAMM parameters are not known beforehand, and they have to be estimated from training data. The estimation algorithm depends on the kind of data available for training. For unsupervised or semi-supervised data (i.e., data whose hidden states are only partly known), all parameters should be estimated simultaneously. In the case of supervised training (i.e., the hidden states of the training data are known), the estimation of the ICAMM parameters can be performed separately for each state and for each chain, and separately from the estimation of the transition probabilities. In the following, we will only consider supervised training.

The transition probabilities $\pi_{\mathbf{kk}'}$ are estimated by counting:

$$\pi_{\mathbf{kk}'} = \frac{\#\text{transitions from combination } \mathbf{k}' \text{ to combination } \mathbf{k}}{\#\text{samples in combination } \mathbf{k}'} \tag{5}$$

This is a maximum likelihood estimator of the transition probabilities, and it is consistent and unbiased. The variance of the estimation depends on the value of $\pi_{\mathbf{kk}'}$, and it is higher for values close to 0.5. The variance of the estimation in this work will be very small, considering that the estimator is consistent and the number of known labels is large (~500 observations). The prior probability of each combination of states is also estimated by counting,

$$p(\mathbf{k}) = \frac{\#\text{samples in combination } \mathbf{k}}{\#\text{samples in any combination}} \tag{6}$$

Since we know the hidden states, the centroids $\mathbf{b}_k^{(l)}$ are estimated empirically. Assuming that N observations are available for training:

$$\mathbf{b}_k^{(l)} = \frac{\sum_{n=1}^{N} \delta_k^{(l)}(n)\, \mathbf{x}^{(l)}(n)}{\sum_{n=1}^{N} \delta_k^{(l)}(n)} \tag{7}$$

where $\delta_k^{(l)}(n)$ is an indicator function whose value is 1 when $c^{(l)}(n) = k$ and 0 otherwise. Given these centroids and the known hidden states, the ICA matrices and the sources of each state and each chain can be calculated using any conventional ICAMM estimation algorithms (e.g., [10, 11]), in such a way that

$$\mathbf{s}_k^{(l)}(n) = \mathbf{W}_k^{(l)}\left(\mathbf{x}^{(l)}(n) - \mathbf{b}_k^{(l)}\right) \forall n / c^{(l)}(n) = k \tag{8}$$

$\mathbf{s}_k^{(l)}(n)$ is only defined for cases where $c^{(l)}(n) = k$, and is undefined elsewhere. As seen in (7) and (8), the

ICA parameters of the *l*th chain (i.e., $\mathbf{W}_{k^{(l)}}^{(l)}, p\left(\mathbf{s}_{k^{(l)}}^{(l)}\right), \mathbf{b}_{k^{(l)}}^{(l)}$, $k^{(l)} = 1,\ldots,K^{(l)}$) are independent from the data and states in other chains, and from the parameters of other states in the same chain. This is due to the decoupling obtained by using supervised training, and a semi-supervised method would have to consider cross-dependencies.

Once the G-SICAMM parameters have been trained, classification can be performed using any method, e.g., by maximum likelihood or the forward-backward method [31]. In this case, we will assume Viterbi decoding, which can obtain the optimal (in the sense of maximum likelihood) solution to the sequence of hidden states [32]. Briefly, Viterbi decoding is a dynamic programming algorithm which searches the most likely sequence of hidden states for a set of *N* observations. This maximization is performed recursively using an auxiliary variable:

$$v_{\mathbf{k}}(n) = p\left(\mathbf{x}(n) \mid \mathbf{c}(n) = \mathbf{k}\right) \cdot \underbrace{\max}_{\mathbf{k'}}\left(\pi_{\mathbf{kk'}} \cdot v_{\mathbf{k'}}(n-1)\right) \tag{9}$$

The values $v_{\mathbf{k}}(n)$ are calculated iteratively, with the initial value being $v_{\mathbf{k}}(1) = p\left(\mathbf{x}(1) \mid \mathbf{c}(1) = \mathbf{k}\right) \cdot p(\mathbf{k})$. Given the proposed G-SICAMM model, $p\left(\mathbf{x}(n) \mid \mathbf{c}(n) = \mathbf{k}\right)$ is calculated as

$$\begin{aligned} p\left(\mathbf{x}(n) \mid \mathbf{c}(n) = \mathbf{k}\right) = \prod_{l=1}^{L} & \left|\det\left(\mathbf{W}_{k_l}^{(l)}\right)\right| \cdot \\ & \cdot \prod_{m=1}^{M^{(l)}} p_s\left(\left(\mathbf{w}_{k_l,m}^{(l)}\right)^T \cdot \left(\mathbf{x}^{(l)}(n) - \mathbf{b}_{k_l}^{(l)}\right)\right) \end{aligned} \tag{10}$$

Once all $v_{\mathbf{k}}(N)$ have been calculated, the algorithm selects the combination of states with maximum $v_{\mathbf{k}}(N)$, $\mathbf{k}_N = \underbrace{\max}_{\mathbf{k}} v_{\mathbf{k}}(N)$, and returns the associated "path" or sequence of states $\mathbf{k}_1,\ldots,\mathbf{k}_{N-1},\mathbf{k}_N$ as the estimated classification of the sequence of observations.

Both the training and the classification algorithms are shown in Table I. The parameters of G-SICAMM are the ICAMM parameters of each chain (i.e., $\mathbf{W}_{k^{(l)}}^{(l)}, p\left(\mathbf{s}_{k^{(l)}}^{(l)}\right), \mathbf{b}_{k^{(l)}}^{(l)}$, $k^{(l)} = 1,\ldots,K^{(l)}$, $l = 1,\ldots,L$), the transition probabilities between every pair of combinations of states, $\pi_{\mathbf{kk'}}$ $\forall \mathbf{k},\mathbf{k'}$, and the priors of each combination of states, $p(\mathbf{k})$.

TABLE I

THE G-SICAMM ALGORITHM.

```
TRAINING
Given a set of data $\mathbf{x}(n)$ and known states $\mathbf{c}(n)$, $n = 1...N_{TRAIN}$
  Estimate the transition probabilities, $\pi_{\mathbf{kk}'}$, using (5)
  Estimate the prior probabilities, $p(\mathbf{k})$, using (6)
  For each chain $l = 1...L$
    For each state $k = 1...K^{(l)}$
      Estimate the centroid $\mathbf{b}_k^{(l)}$ using (7)
      Jointly estimate the demixing matrices $\mathbf{W}_k^{(l)}$ and the sources $\mathbf{s}_k^{(l)}$ using an embedded ICA method, e.g., [10, 11]
CLASSIFICATION
Given a set of data $\mathbf{x}(n)$, $n = 1...N_{TEST}$
  Estimate $p(\mathbf{x}(1) \mid \mathbf{c}(1) = \mathbf{k})$ for all $\mathbf{k}$ using (10)
  Initialize $\nu_{\mathbf{k}}(1) = p(\mathbf{x}(1) \mid \mathbf{c}(1) = \mathbf{k}) \cdot p(\mathbf{k})$
  For each observation $n = 1...N_{TEST} - 1$
    Estimate $p(\mathbf{x}(n) \mid \mathbf{c}(n) = \mathbf{k})$ for all $\mathbf{k}$ using (10)
    Update $\nu_{\mathbf{k}}(n)$ using (9)
  Estimate $\mathbf{k}_{N_{TEST}}$ by maximizing $\nu_{\mathbf{k}}(N_{TEST})$
  For each observation $n = N_{TEST} - 1...1$
    Select the observation that follows the path that ends in $\mathbf{k}_{N_{TEST}}$
```

IV. SIMULATIONS

The classification performance of G-SICAMM was measured by several Monte Carlo experiments on simulated data. The following methods were considered for comparison: GMM; ICA mixture models (ICAMM); continuous HMM whose emissions were modeled using GMM (HMM); continuous HMM whose emissions were modeled using ICA (SICAMM); continuous coupled HMM whose emissions were modeled using GMM (CHMM); long short-term memory (LSTM) recurrent neural networks; and the proposed method, G-SICAMM. We considered three simulated experiments: (i) determining the behavior of G-SICAMM with respect to time dependencies within the same chain (Section IV.C.1); (ii) determining the behavior of G-SICAMM with respect to time dependencies between different chains (Section IV.C.2); and (iii) a sensitivity analysis to test the behavior of G-SICAMM with respect to the number of chains, L (Section IV.D).

A. *Model parameter initialization*

The simulated data were obtained from a G-SICAMM with two chains $(L = 2)$, two hidden states in each chain $\left(K^{(1)} = K^{(2)} = 2\right)$, and observations of dimension four in both chains $\left(M^{(1)} = M^{(2)} = 4\right)$. The parameters of the model were initialized as follows:

- The demixing matrices for each state and each chain, $\mathbf{W}_{k_l}^{(l)}$, were randomly initialized using values drawn from a uniform distribution in the range [0, 1].
- The centroids were set relatively close, $\mathbf{b}_1^{(1)} = \mathbf{b}_1^{(2)} = [1,1,1,1]^T$ and $\mathbf{b}_2^{(1)} = \mathbf{b}_2^{(2)} = [1.5,1.5,1.5,1.5]^T$.
- The sources followed a non-Gaussian (uniform) distribution with zero mean and unit standard deviation.

- In order to set the amount of time dependence in the data, the transition probabilities of both chains were obtained using the following conditional transition probabilities:

$$\begin{array}{l}
\text{If the other chain was in state 1 at time } n-1 \\
\pi_{11}^{(l)}(n) = \pi_{22}^{(l)}(n) = \alpha \\
\pi_{12}^{(l)}(n) = \pi_{21}^{(l)}(n) = 1-\alpha \\
\text{If the other chain was in state 2 at time } n-1 \\
\pi_{11}^{(l)}(n) = \pi_{22}^{(l)}(n) = \alpha-\beta \\
\pi_{12}^{(l)}(n) = \pi_{21}^{(l)}(n) = 1-(\alpha-\beta)
\end{array} \qquad (11)$$

where $0 \le \alpha \le 1$ and $\alpha - 1 \le \beta \le \alpha$, $l = 1,2$. α and β are two parameters which allow us to regulate the sequential dependence in the data. α is the intra-chain dependence parameter, since it sets the time dependence of each chain with respect to past values of the same chain. Conversely, β is the inter-chain dependence parameter, since it sets the time dependence of each chain with respect to past values of the other chain. If $\beta = 0$, the transition probabilities of each chain are independent from those of the other chain; furthermore, if $\beta = 0,\ \alpha = 0.5$ there is no time dependence. The transition probabilities between combinations of states, $\pi_{\mathbf{k}\,\mathbf{k}'}$, are calculated from those in (11) using results from probability theory.

These values were selected in order to obtain a model that was simple, yet informative enough to show the behavior of the proposed method. The resulting G-SICAMM model, however, is flexible enough to be used in many real applications. For instance, the number of chains and states were consistent with those used in the experiment on real data (Section V), where the two states corresponded to a binary classification (active/inactive) of the brain regions delimited by the chains.

B. Data generation procedure

Once defined the G-SICAMM parameters, we can randomly generate data for this model configuration using the following procedure. Let us assume that we need to generate *N* data samples. First, we assume that the model starts at the combination of states $\mathbf{k} = [1,\ 1]^T$ and use the transition probabilities $\pi_{\mathbf{k}\,\mathbf{k}'}$ to generate a random sequence of *N* combinations of states. Once the states have been generated, the observations can be obtained as explained in Table II.

TABLE II
DATA GENERATION ALGORITHM FROM G-SICAMM ONCE THE STATES ARE KNOWN.

For each chain, $l = 1 \ldots L$:
For each time instant, $n = 1 \ldots. N$:
For each source, $m = 1 \ldots M^{(l)}$:
Draw one random value from the distribution of the *m*th source of state $c^{(l)}(n) = k^{(l)}$
Multiply the sources by the mixing matrix of the corresponding state, $\left(\mathbf{W}_{k^{(l)}}^{(l)}\right)^{-1}$, and add the centroid, $\mathbf{b}_{k^{(l)}}^{(l)}$

C. Simulated Experiments

For each iteration of the experiment, a known G-SICAMM model was set as indicated in Section IV.A and $N = 1024$ observations, $\mathbf{x}(n),\ n = 1 \ldots N$, along with their respective combinations of states, $\mathbf{c}(n)$, were randomly drawn from the model using the method outlined in Section IV.B. The first half of the data was used for training and the second half of the data was used for testing the classification performance of every competitive method. Performance was measured using the average classification error rate for 300 iterations.

The considered methods were set up as follows. For the coupled HMM methods (i.e., CHMM and G-SICAMM), the models were set as per the simulation, with two chains $\left(L=2\right)$, each one with two hidden states $\left(K^{(1)}=K^{(2)}=2\right)$ and with the emissions being vectors of size 4 $\left(M^{(1)}=M^{(2)}=4\right)$. For the single-chain methods (HMM, SICAMM, GMM, ICAMM, and LSTM), both chains were considered at the same time, resulting in four hidden states (the combinations of states of each chain, $K'=2{\cdot}2=4$) and emissions of size eight $\left(M'=4+4=8\right)$. For G-SICAMM, SICAMM and ICAMM, the ICA for each hidden state was estimated from training data, assuming square mixing matrices, using MIXCA [11].

For GMM, HMM and CHMM, the number of Gaussian components for each hidden state ranged between 3 and 20 components, a range of numbers used in several EEG applications, e.g., [33]). The exact number of components for each state was chosen by maximization of the Akaike information criterion [34], a commonly-used criterion to test the quality of a statistical model. There is extensive literature of the application of AIC criterion for choosing the optimal number of components, including very recent works (e.g., [35, 36]).

In all cases, prior probabilities and transition probabilities were estimated by counting.

The parameters of the LSTM network were chosen using cross-validation on the training set. For the simulated case, we opted with a network with six layers: an input layer, a fully connected layer with 50 neurons, a ReLu layer, a bidirectional LSTM layer with 6 neurons, a second fully connected layer with one neuron per combination of states (K'), and a softmax layer. The network was trained using Adam with starting learning rate 0.01.

1) Intra-chain time dependence

For the first experiment, the inter-chain dependence parameter (11) was set to $\beta=0.1$ and the intra-chain dependence parameter was changed from $\alpha=0.5$ (no dependence) to $\alpha=0.99$ (almost complete dependence) in steps of 0.025. The Monte Carlo experiment was repeated 300 times for each value of α, for a total number of 6,000 iterations.

Fig. 2 shows the average classification rate of each method. For readability, no variance bars are shown. However, the variance of the classification rate was low overall, lowered with inter-chain dependence, and was lower for ICA-based methods than it was for LSTM and GMM-based methods. The non-dynamic methods (ICAMM and GMM) maintained their performance when the intra-chain dependence increased. Conversely, all the dynamic methods (LSTM, HMM, SICAMM and G-SICAMM) increased their performance as α increased. SICAMM performed better than HMM and very similarly to LSTM and CHMM. Finally, G-SICAMM achieved the best classification performance at every point of the simulation due to exploitation of time cross-dependencies between chains.

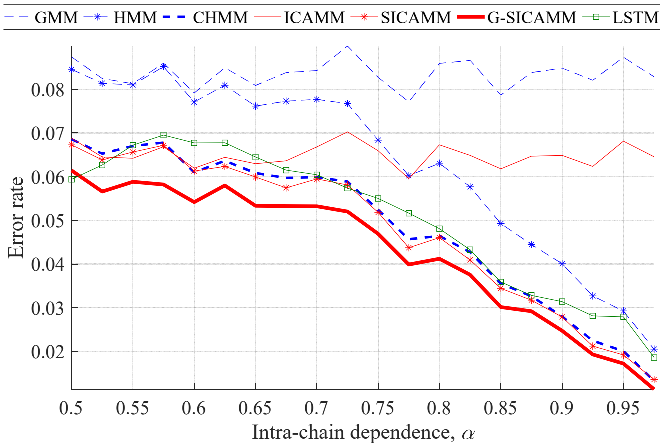


Fig. 2. Average classification error rate with respect to the intra-chain dependence parameter, α. For readability, no variance bars are shown: variance was always small and reduced with increasing α.

2) *Inter-chain time dependence*

A second experiment was carried out to test the performance of the methods with respect to changes in the inter-chain dependence parameter (11). In this case, the intra-chain dependence parameter was set to $\alpha = 0.8$ and the inter-chain dependence parameter was changed from $\beta = 0$ (no dependence) to $\beta = \alpha = 0.8$ (maximum dependence) in steps of 0.06. This relatively high value of α was set in order to allow for a larger range of variation for β, since $\beta \leq \alpha$. The Monte Carlo experiment was repeated 300 times for each value of β, for a total number of 3,900 iterations.

Fig. 3 shows the average classification rate of each method. As in Fig. 2, no variance bars are shown, since variance was low overall, and lower for ICA-based methods than it was for LSTM and GMM-based methods. Also in concordance with the results shown in Fig. 2, G-SICAMM consistently outperformed the other competitive methods, with SICAMM and CHMM tied for the second best result. In this experiment, LSTM yielded an intermediate result between SICAMM/CHMM and HMM. HMM yielded a similar result to the non-dynamic ICAMM for some range of values of β, but yielded an overall better result than ICAMM. Finally, ICAMM and GMM achieved steady results since they do not consider dependence.

In Fig. 3, the performance of dynamic methods (G-SICAMM, SICAMM, LSTM, CHMM and HMM) worsened with rising inter-chain dependence $\beta < 0.2$, and only improved when $\beta > 0.3$. This behavior can be explained from the transition probabilities set for the model in (11). Since $\alpha = 0.8$, positive values of β will actually reduce the time dependence in the model as long as $\alpha - \beta \geq 0.5$. If we keep increasing the inter-chain dependence parameter so that $\alpha - \beta < 0.5$, the overall time dependence in the model rises again. This explains the results in Fig. 3, where the effect of the time dependence is more pronounced the more β steps away from 0.3 (i.e., the more $\alpha - \beta$ steps away from 0.5). Regardless of this effect, G-SICAMM obtained a better result than the other considered methods for any value of inter-chain dependence.

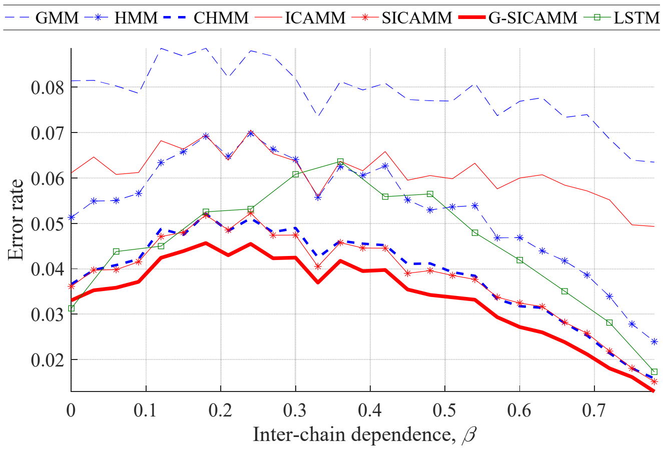


Fig. 3. Average classification error rate with respect to the inter-chain dependence parameter, β. For readability, no variance bars are shown: variance was always small and reduced with increasing β.

Please notice that every point in the curves of Figures 2 and 3 corresponds to a specific dependence model determined by α and β, so it is normal that G-SICAMM and SICAMM follow similar error patterns. This is because in a given point, both methods must estimate the same dependence model. However, G-SICAMM needs to estimate just two parameters (α, β) using the sample estimates of equation (5) to compute the transition matrices of equation (11). On the contrary, SICAMM, as explained before, fusions both chains in only one and so defines four states instead of two. Thus, 10 parameters must be estimated from equation (5) corresponding to the symmetric (4x4) transition matrix. For larger dimensions, the difference between the transition parameters required by SICAMM and G-SICAMM increases dramatically. Moreover, G-SICAMM requires the estimation of two ICAMM models of dimension 4, while SICAMM must estimate one model of dimension 8. This also contributes to a constant (independent of α and β) excess error of SICAMM respect to G-SICAMM. All this explains the larger error showed by SICAMM and ultimately justifies the improved performance of G-SICAMM in both the simulated and the real examples, as it will be shown in Section V.

D. Sensitivity analysis

A further experiment was set up to show the effect of the number of chains on the performance of the model. In previous sections, the parameters of G-SICAMM were initialized in a supervised learning scheme considering that the number of chains was known. The experiment in this section was designed in order to approximate a more general case where the number of chains is unknown. The parameters of the generating model for this experiment were set as explained in Section IV.A, using the following values for the parameters of the model:

- Four chains $(L=4)$, each with four variables $\left(M^{(l)}=4,\ l=1...L\right)$ and two hidden states $\left(K^{(l)}=2,\ l=1...L\right)$. Thus, in total, each observation $\mathbf{x}(n)$ had 16 variables.
- The centroids of the model were $\mathbf{b}_1^{(l)}=[1,1,1,1]^T$ and $\mathbf{b}_2^{(l)}=[1.5,1.5,1.5,1.5]^T$ for all chains, $l=1...L$.
- The sources followed a non-Gaussian (uniform) distribution with zero mean and unit standard deviation.
- In order to simplify dependence, only the following 4 combinations of states were considered: $\mathbf{k}_1=[2,1,1,1]^T$, $\mathbf{k}_2=[1,2,1,1]^T$, $\mathbf{k}_3=[1,1,2,1]^T$ and $\mathbf{k}_4=[1,1,1,2]^T$. This guaranteed that all chains crossed

their two states, while keeping the number of combinations to a minimum. The resulting $\left[4\times 4\right]$ transition matrix was set to $\pi_{\mathbf{kk}}=0.6$ and $\pi_{\mathbf{kk}'}=0.13$ for $\mathbf{k}'\neq\mathbf{k}$.

$N=1024$ data were generated from this model as explained in Section IV.B. The first half of this data was used to fit G-SICAMM models with a different number of chains, ranging from 1 to 8. For models with a smaller number of chains than the generating model, we proceeded as for SICAMM in Section IV.C (i.e., by combining the data and establishing the possible combinations of states). For models with a higher number of chains than the generating model, one or more chains were split in half, obtaining two chains with observations of size 2 and with the same hidden states as the original chain. These parameters are summarized in Table III.

The second half of the generated data was used for testing the models. This process was repeated for 300 iterations, and performance was measured using the average classification error rate. Fig. 4 shows the results of the simulation. There is a clear optimum for the correct number of chains ($L=4$). This shows the sensitivity of the results with respect to the number of chains, which requires enough flexibility to model the original data. Furthermore, smaller numbers of chains (1-3) obtained better performance than higher numbers of chains (5-8). This could be due to the reduced complexity of the smallest models, which could offset the incorrect number of chains. Therefore, in case of similar or conflicting results, one should likely choose the lowest number of chains.

TABLE III
PARAMETERS OF EACH G-SICAMM USED FOR THE SENSITIVITY ANALYSIS EXPERIMENT.

Number of chains			
1	**2**	**3**	**4**
$M^{(1)}=16$	$M^{(1)}=8$ $M^{(2)}=8$	$M^{(1)}=8$ $M^{(2)}=M^{(3)}=4$	$M^{(l)}=4\forall l$
$K^{(1)}=4$	$K^{(1)}=4$ $K^{(2)}=4$	$K^{(1)}=4$ $K^{(2)}=K^{(3)}=2$	$K^{(l)}=2\forall l$
Number of chains			
5	**6**	**7**	**8**
$M^{(l)}=4, l=1\ldots3$ $M^{(4)}=M^{(5)}=2$	$M^{(1)}=M^{(2)}=4$ $M^{(l)}=2, l=3\ldots6$	$M^{(1)}=4$ $M^{(l)}=2, l=3\ldots7$	$M^{(l)}=2\forall l$
$K^{(l)}=2\forall l$	$K^{(l)}=2\forall l$	$K^{(l)}=2\forall l$	$K^{(l)}=2\forall l$

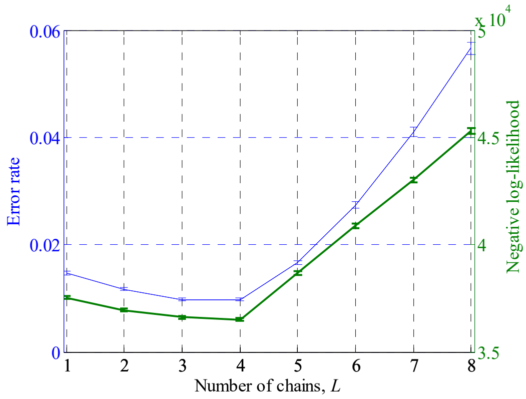


Fig. 4. Results obtained by changing the number of chains in which we divide the data: a) error rate; b) negative log-likelihood. The number of chains of the generative model was 4.

V. Real Application

One of the main fields of application of pattern recognition is bioinformatics, which includes the analysis of electroencephalographic (EEG) data [37]. EEG signals are a useful clinical tool and they are an active area of research, for instance, on cognitive functions [38], sleep disorder diagnosis [39], and brain connectivity analysis [40]. Previous works have shown that ICA can produce not only a valid statistical model of the EEG data, but also sources with physiological significance. Typically, each column of the mixing matrix is considered as a spatial pattern ("scalp map") of activation of the sources, with the corresponding source being interpreted as the level of activation of the map during the experiment. Popular applications of ICA on EEG have been to locate dipolar sources in the brain and to remove artifact (non-EEG) sources (see [24, 26] and the references within). HMM has also been applied to analyze the event-related dynamics of brain oscillations and the causality of physiological phenomena [37]. Two examples of this application are decoding upper limb movement EEG signals from chronic stroke patients with impaired mobility [41] and controlling brain computer interfaces [42]. In general, dynamic analyses have assumed GMM-based methods that provide adequate statistical modeling capabilities, but constrain a physical interpretation or association with the underlying grounds of the physiological phenomenon.

In this paper, we have approached the processing of EEG signals from epileptic patients that were performing neuropsychological tests. These tests were applied as part of clinical diagnosis analyses to monitor the learning and memory cognitive function of the patients. In particular, we were interested in studying changes in the activation of brain areas during memory tasks. This kind of tests makes up an essential area of clinical neurophysiology assessment. Information cannot be processed if the brain is unable to store a certain amount of it in short-term (working) memory or to consolidate past experiences, events and strategies in long-term memory. Conversely, information stored in short- or long-term memory is useless without the means to properly access and activate it.

In patients with refractory temporal lobe epilepsy, it is important to assess hemispheric memory and speech lateralization because preexisting memory deficits may worsen or new deficits may appear after surgical resection. The standard preoperative assessment of the lateralization of memory and speech includes intracarotid amytal testing, which is an invasive procedure. However, the (non-invasive) hemispheric EEG analysis using the proposed G-SICAMM method during the application of neuropsychological memory tests may be a useful tool to evaluate the memory function dominance. There is clinical evidence that the neural networks of each hemisphere show independent activation patterns following different dynamics. Collaborative interaction between hemispheres, however, is required to accomplish some cognitive tasks. From a medical standpoint, two areas (left and right cerebral hemispheres) may be a large enough number of areas to be analyzed. Therefore, a G-SICAMM with two Markov chains, one for each hemisphere, was considered for this application. In other EEG applications, like spindle sleep analysis, the number of areas to be analyzed may be four or more [43].

A. Experimental Setup

The number of epileptic patients was six. The biosignal acquisition and the tests were synchronized using a graphic user interface designed by the authors, and the data were captured by the Neurophysiology and Neurology Units at Hospital La Fe, Valencia.

The first neuropsychological test was the visual memory Barcelona Neuropsychological Test (TB, [21]). The test consists of ten trials and scoring is given depending on the number of correct responses. In each trial, the participant is shown an abstract line figure (probe) during 3 seconds. There is a 2-second retention interval and afterward the participant is told to recognize the probe out of a group of four similar figures. The second test was the TAVEC [22], an auditory working memory test. In each trial, the participant listens to a list of sixteen items and is then told to repeat as many items as they can remember.

The stages of the TB can be split in two states ([stimulus+retention] vs. [response]) or three states ([stimulus] vs. [retention] vs. [response]). The stages of the TAVEC can be split in two states ([stimulus] vs.

[response]). The proposed methods were used for classifying the EEG signals into the chosen number of states. The first half of the data of each patient was used for training, and the second half of the data was used for testing.

For each subject, 19 EEG channels were captured using a sampling frequency of 500 Hz and positioned according to the 10-20 system. The signals were filtered and split into epochs of 0.25 second length. This short length was selected in order to ensure that all stages of the test (some of which were very short) were spread over multiple epochs, thus improving parameter estimation. The twelve features shown in Table IV were calculated for each epoch, and the best feature subset for each subject was selected using cross-validation on the training set. Some of these features (TSI, ASI, spindles, average amplitude, centroid frequency) are commonly used in polysomnography analysis [14, 44]. The statistical features were based on higher order statistics, which have been used in several applications on EEG data (see [45] and the references within). The remaining features are classical and have been used in many works.

We considered the same methods used on simulated data. The parameters of the ICAMM, SICAMM and G-SICAMM methods were estimated using supervised training on the first half of the data, following the method described in Section IV.C, except that $M'=19$. G-SICAMM and CHMM were set up with two chains $\left(L=2\right)$, one for each brain hemisphere $\left(M^{(1)}=M^{(2)}=9\right)$. Both chains considered the same states: in this case, the chain structure is used to isolate the contributions of each brain hemisphere. Such division could also be used, for instance, in order to determine hemispheric dominance during certain tasks and measure spatial neglect ([46]). The parameters of the LSTM network were again chosen using cross-validation on the training set. The results shown in the following correspond to a network with six layers: an input layer, a fully connected layer with 100 neurons, a ReLu layer, a bidirectional LSTM layer with 6 neurons, a second fully connected layer with one neuron per combination of states (K'), and a softmax layer. The network was trained using Adam with starting learning rate 0.01.

The performance of each method was initially measured using the error rate and the recall (or sensitivity) of the classification. However, the stages of the test have different durations, with the stimulus stage being the longest. Thus, the error rate and the recall were heavily dominated by this stage. In order to compensate for these differences in prior probability, we replaced the error rate with the balanced error rate (BER) and the recall with Cohen's kappa coefficient (κ, [23]). The BER is the average of the error rates for each state, and thus is more resistant with respect to states with very different prior probabilities. Cohen's kappa coefficient is a commonly-used tool to assess accuracy that takes into account different state probabilities, and thus it is much more robust than the recall. In this work, the following definition of κ for a classification with *K* states was used

$$\kappa=\frac{N\cdot\sum_{i=1}^{K}c_{ii}-\sum_{i=1}^{K}c_{i+}\cdot c_{+i}}{N^2-\sum_{i=1}^{K}c_{i+}\cdot c_{+i}} \tag{12}$$

where *N* is the number of classified samples; c_{ii} is the number of correctly classified samples from state *i*, $i=1,...,K$; c_{i+} is the number of samples that truly belong to state *i*; and c_{+i} is the number of samples that were classified as state *i*. The BER is bound in the interval [0, 1], and a low value is better than a high value. Cohen's kappa is bound in the interval [-1, 1] and a high value is better than a low value. Thus, an optimum result would have BER = 0 and $\kappa=1$.

TABLE IV
LIST OF FEATURES CALCULATED FROM THE DATA x AT EACH EPOCH OF LENGTH Δ.

Feature	Definition

Feature	**Definition**
Average amplitude	$y_1(n)=\frac{1}{\Delta}\sum_{i=n}^{n+\Delta}\lvert x(i)\rvert$
Maximum amplitude	$y_2(n)=\max\{\lvert x(n)\rvert,\ldots,\lvert x(n+\Delta)\rvert\}$
Average power	$y_3(n)=\frac{1}{\Delta}\sum_{i=n}^{n+\Delta}\left(x(i)-y_1(n)\right)^2$
Centroid frequency	$y_4(n)$ = pole frequency of AR2 model
Peak frequency	$y_5(n)=\underbrace{\max}_{f}\lvert X(f)\rvert^2$, where $X(f)$ is the Fourier transform of $\left[x(n),\ldots x(n+\Delta)\right]$
Spindles ratio	$y_6(n)=\frac{P_{sigma}}{P-P_{sigma}}$, where $P=\int\lvert X(f)\rvert^2\cdot df$ and $P_{sigma}=\int_{11.5Hz}^{15Hz}\lvert X(f)\rvert^2\cdot df$
TSI	$y_7(n)=\frac{P_{theta}}{P_{delta}+P_{alpha}}$, where $P_{delta}=\int_{0.5Hz}^{3.5Hz}\lvert X(f)\rvert^2\cdot df$, $P_{theta}=\int_{3.5Hz}^{8Hz}\lvert X(f)\rvert^2\cdot df$, and $P_{alpha}=\int_{8Hz}^{11Hz}\lvert X(f)\rvert^2\cdot df$
ASI	$\mathrm{y}_8(n)=\frac{P_{alpha}}{P_{delta}+P_{theta}}$
Skewness	$y_9(n)=\frac{1}{\Delta}\sum_{i=n}^{n+\Delta}\frac{\left(x(i)-y_1(n)\right)^3}{\left(y_3(n)\right)^{3/2}}$
Kurtosis	$y_{10}(n)=\left(\frac{1}{\Delta}\sum_{i=n}^{n+\Delta}\frac{\left(x_i(n)-y_1(n)\right)^4}{\left(y_3(n)\right)^2}\right)-3$
Time reversibility	$y_{11}(n)=\left(y_3(n)\right)^{-3/2}\cdot\frac{1}{\Delta-\tau}\cdot\sum_{i=n+\tau}^{n+\Delta}\left[x(i)-x(i-\tau)\right]^3$, with $\tau=1$
Third-order autocovariance	$y_{12}(n)=\frac{1}{\Delta-2\tau}\sum_{i=n+2\tau}^{n+\Delta}x(i)\cdot x(i-\tau)\cdot x(i-2\tau)$, with $\tau=1$

B. Results

Fig. 5.a shows the average BER and kappa values of each method for two-class classification ([stimulus+retention] vs. [response]). In concordance with the results in Section IV, ICAMM-based models obtained a better overall performance than GMM-based models, and dynamic methods performed better than non-dynamic methods. The latter is particularly marked in the case of G-SICAMM, with considerably better results (lower BER and higher κ) than ICAMM. G-SICAMM obtained the best results overall, and SICAMM and LSTM yielded worse performances than G-SICAMM. In all cases, however, it can be seen that this classification is a difficult problem that required the use of powerful dynamic models. These trends remain consistent when considering three-class classification ([stimulus] vs. [retention] vs. [response]), as seen in Fig. 5.b), although with slightly higher BER and lower Cohen's kappa for all methods. Once more, dynamic methods outperformed non-dynamic methods, and G-SICAMM obtained the best results. Similar trends were obtained for two class classification for the TAVEC test, as shown in Fig. 5.c, with the exception of LSTM, which yielded a result comparable to that of G-SICAMM in this case.

Fig. 6 shows the results of the classification of the two neuropsychological tests for one of the subjects: Barcelona test with two states (Fig. 6.a) and three states (Fig. 6.b), and TAVEC test with two states (Fig. 6.c). Results for other subjects were similar to those presented. The results of the non-dynamic methods (GMM, ICAMM) tended to oscillate very fast or to remain stuck at one particular class, thus explaining the worse performance yielded by those methods. CHMM and HMM achieved good (e.g., Fig. 6.a, .c) and bad (e.g., Fig. 6.b) results in several trials, with an overall better result than both non-dynamic methods. The dynamic ICAMM-based methods and LSTM consistently yielded the best results. LSTM and SICAMM showed some oscillating behavior, while G-SICAMM showed very few rapid changes, yielding very smooth classifications. This behavior produced better average performance for G-SICAMM, although it still fails in some trials (see G-SICAMM in Fig. 6.c at the 720s mark). This is also supported by the high values of kappa, which seems to indicate high concordance between the true states and the classification obtained by the proposed method.

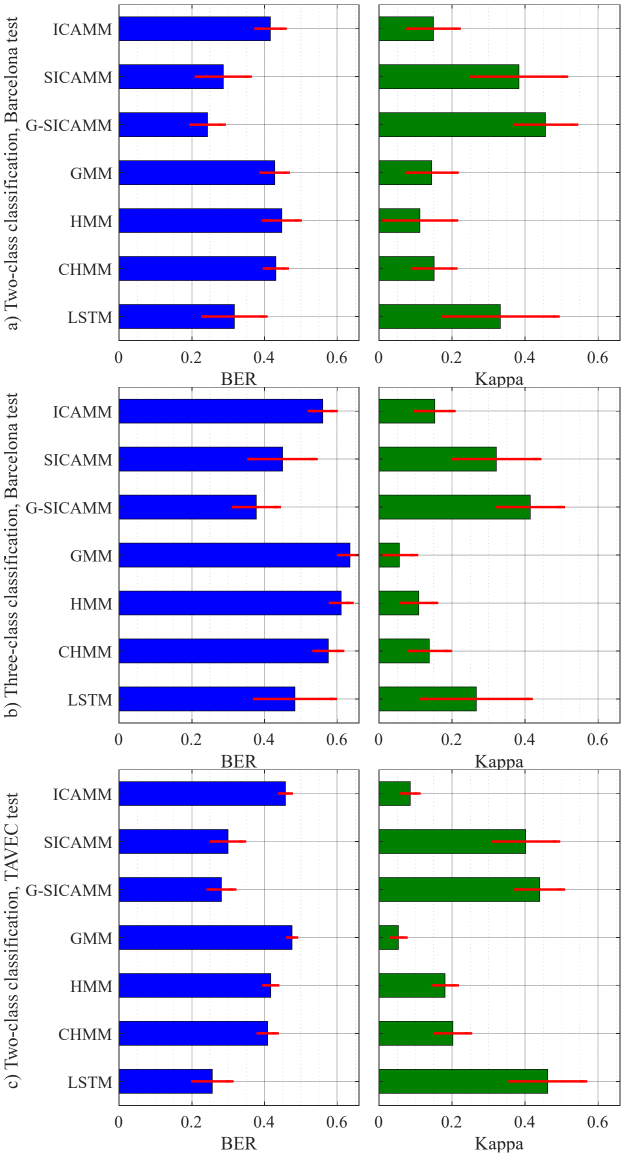


Fig. 5. Average results of the classification of multimodal data from patients performing neuropsychological tests, in terms of balanced error rate and Cohen's kappa. Error bars are shown in red.

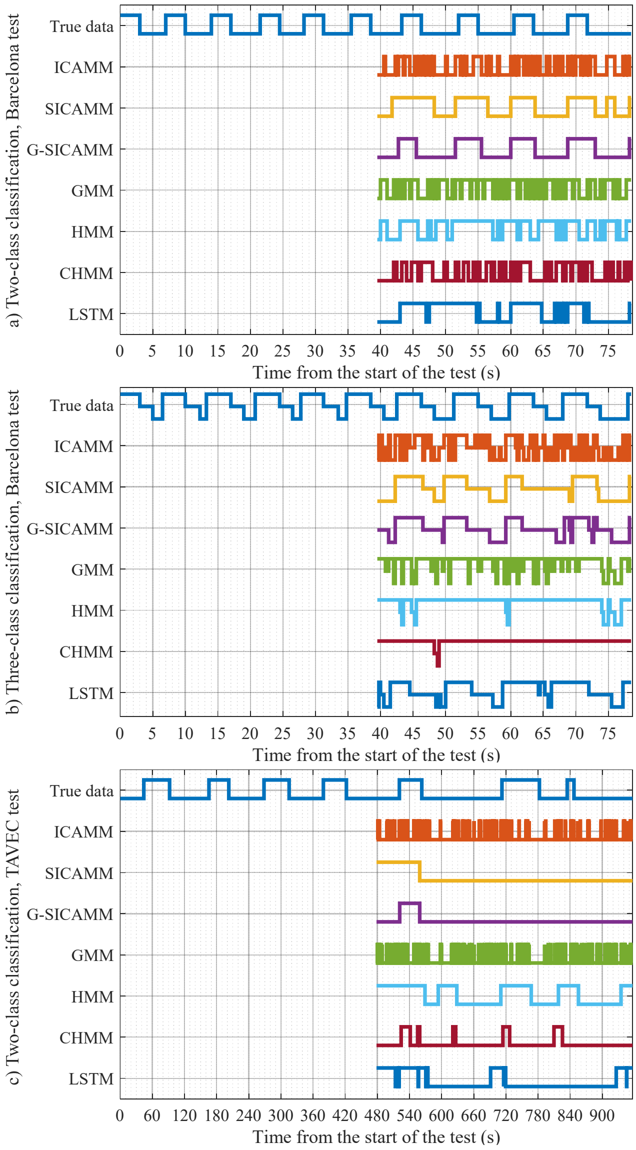


Fig. 6. Classification of neuropsychological tests for one of the subjects.

As mentioned in Section III, the parameters of G-SICAMM can be used to model the dynamics of EEG signals in a similar way to ICAMM parameters [24]. In brief, the mixing matrices are considered as a set of spatial patterns (or "scalp maps") whose independent temporal activations are regulated by their corresponding sources. The study of these parameters can determine the connectivity of brain regions during the stages of the experiment and shed new light on the EEG data (e.g., [47]). Conversely, the parameters obtained by the Gaussian mixed model have no direct relation with the physiological parameters underlying the EEG data. Thus, the hidden Markov models with GMM emissions can only be used for classification, while G-SICAMM can be used both for classification and for understanding the physiological processes behind the EEG. Similarly, the weights of the LSTM network are not readily interpretable as physiological phenomena, although some works have attempted to project results back onto the scalp for convolutional neural networks [48].

Some of these preliminary results are shown in Fig. 7. Fig. 7.a shows some results of the Barcelona test for subject #3, and Fig. 7.b shows some results of the TAVEC test for the same subject. In both cases, the independent components found during the stages of the experiment were consistent with the nature of the experiment itself. Furthermore, there were differences between the sources of each state, whereas sources from the same state were more similar between them. For the Barcelona test, the results show the activation of two regions typical in visual tasks. The first one is the excitation of the occipital region of the head, which is related to visual input and processing. Thus, its presence is in concordance with the test. The other common source is distributed on the frontal-central region of the head. This region has been found to be related with the processing of information during working memory, as seen in [49]. Since this experiment is based on working memory, it is indeed consistent that such an area would appear consistently across experiments. Finally, the third interesting source found in the response events, which seems to be more noisy than the other two, might be caused by electromyographic noise due to the movement required during the response (hand movement and pointing).

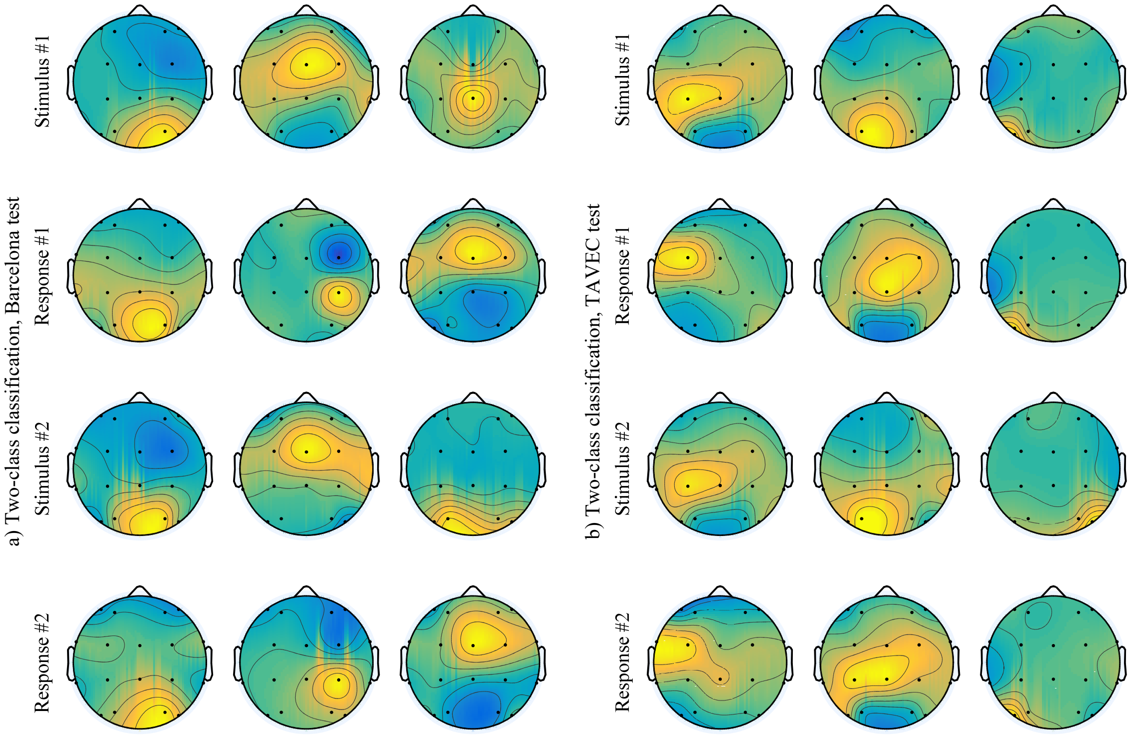


Fig. 7. Spatial patterns of three of the sources extracted from subject #3: a) Barcelona test; b) TAVEC test.

For the TAVEC test (Fig. 7.b), the patterns for the stimulus state show higher responses on the left hemisphere, particularly on the temporal area. This suggests the activation of Wernicke's speech area on the dominant (left) cerebral hemisphere [50], which is in concordance with the fact that the TAVEC is a verbal test. The activation of the occipital cortex might suggest visual activity on the part of the subject. Besides, the activity during the response state is more centered and more oriented towards the front area of the brain. This suggests a suppression of the visual input during the responses, which fits the nature of the test. Other interesting result reflected in the estimated G-SICAMM parameters is the excitation of the left frontal temporal area which suggests the activation of Broca's speech area, usually related with the production of speech [50].

VI. Discussion and Conclusion

The method proposed, G-SICAMM, defines a general pattern recognition framework to characterize the joint behavior of a number of synchronized SICAMM models. Subspace classification, learning of interaction dynamics, and hidden non-Gaussian source separation are jointly considered. Thus, the proposed method is capable of modeling several datasets in parallel, learning local dynamic accurately while also considering global dynamic interactions. The requirements of the problem (physical components, sensor spatial locations, data modalities,...) would determine the use of several Markov chains. Thus, given the characteristics of a problem, simpler methods might be used, e.g., SICAMM, if there are no physical, topological or functional reasons to use several chains and ICAMM if there is no temporal dependence.

Some of the parameters of G-SICAMM are fixed due to the method design, e.g., since the mixing matrices are square, the number of sources per chain is equal to the number of variables per chain. In the proposed supervised learning scheme, the number of states and the number of chains (which depend on the characteristics of the application) are considered as prior information. Classification without prior information would require an unsupervised learning scheme for G-SICAMM that is outside the scope of this work. However, the restriction about the number of chains would remain conditioned to the application.

The inner structure of the observed data in the multichannel setting makes G-SICAMM a more efficient method to deal with a limited amount of data for both training and testing in comparison with one channel methods like SICAMM or HMM. It is also more versatile than other coupled methods like CHMM, as non-Gaussian modelling is possible. Besides, the inherent ICA structure of G-SICAMM may lead to a better interpretation of the results by analyzing the estimated hidden independent sources. Simulations showed that the higher the value of temporal dependency in HMM (intra- and inter-chain), the higher the improvement in classification error rate of G-SICAMM over those of the other competitive methods. These characteristics of G-SICAMM would be suitable for the analysis of brain networks in order to model relationships of activation/inhibition of relatively distant zones.

The flexibility of G-SICAMM to model data densities with different degrees of alteration, changing in time, of an underlying ICA mixture model was also demonstrated. Accurate non-Gaussian source approximation depends on the embedded ICA algorithm and the way that the multichannel dataset is divided. The first depends on the assumptions of the ICA estimation, being non-parametric the most relaxed, and the latter normally is related with the application. Thus, the configuration of G-SICAMM parameters should be tuned with the application requirements. In this work, G-SICAMM was able to adapt to different cases of changing non-stationarity and non-linearity from EEG signals.

Mixture models (MM) have been employed in many applications due to their suitability to represent complex data geometries. One of the most popular MM is the Gaussian Mixture Model (GMM), whose data generating process has been successfully adapted for description and prediction in several applications (see for instance [51]). In practice, an adequate deployment of GMM is able to model almost any kind of data distribution. However, the capabilities of inference from the parameters of GMMs are constrained since it is difficult to associate the estimated Gaussian components with variables from real applications. Thus, the reasoning of assigning meaningful patterns to GMM components based on hypotheses on a real phenomenon is not achievable. In contrast, non-Gaussian mixture models (NGMM) would enable the estimation of plausible underlying random variables that could be related with physical phenomena, besides the usefulness of generating a predictive distribution. This was demonstrated in the processing of EEG signals.

We considered the automatic classification of EEG signals from epileptic patients performing two learning and memory neuropsychological tests. Two types of classification were considered: two-class classification ([stimulus+retention] vs. [response]) and three-class classification ([stimulus] vs. [retention] vs. [response]). G-SICAMM outperformed all the other competitive methods, including a continuous two-coupled HMM and a LSTM recurrent network. The average improvements obtained by G-SICAMM over LSTM and GMM-based CHMM were 0.17 and 0.08 (for the balanced error rate) and 0.27 and 0.05 (for Cohen's kappa coefficient), respectively. The configuration of G-SICAMM allowed brain hemisphere analysis, and thus, more accurate detection of small dynamic changes in each hemisphere than the one obtained by analyzing the brain dynamics as a whole.

Furthermore, the G-SICAMM parameters provided a structured result from which an explanation from the application standpoint was made. The independent components found for the six patients during the stages of the neuropsychological test were consistent with the underlying physiological model. For the visual test, there were activations of spatial zones of the brain dedicated to the processing of visual information and attention, and of the premotor cortex. For the auditory test, we confirmed the activation of brain regions closely related with understanding and producing speech. Thus, the potential of the method for the analysis of brain dynamics on EEG signals was demonstrated.

There are several lines of work that remain open to exploration: (i) to study in greater depth the dynamic model parameters extracted from real biological data and search for any correlation with the underlying biological processes (e.g., EEG connectivity); (ii) to incorporate different kind of enhancements such as semi-blind source separation; unsupervised and semi-supervised learning; and multimodal pattern modeling; (iii) to adapt the treatment of the CHMM structure to consider alternatives to the current fully connected approach, which would reduce computational time and improve numerical convergence; and (iv) to introduce G-SICAMM in different applications such as multimodal fusion.

ACKNOWLEDGMENT

This work was supported by Spanish Administration (Ministerio de Economía y Competitividad) and European Union (FEDER) under grants TEC2014-58438-R and TEC2017-84743-P.

REFERENCES

[1] O. Cappe, E. Moulines, and T. Ryden, *Inference in Hidden Markov Models*. New York, NY: Springer, 2005.

[2] L. Liu, L. Shao, F. Zheng, and X. Li, “Realistic action recognition via sparsely-constructed Gaussian processes,” *Pattern Recognition*, vol. 47, no. 12, pp. 3819-3827, 2014.

[3] A.L. Tambo, B. Bhanu, N. Ung, N. Thakoor, N. Luo, and Z. Yang, "Understanding pollen tube growth dynamics using the Unscented Kalman Filter," *Pattern Recognition Letters*, vol. 72, pp. 100-108, 2016.

[4] K.S. Xu and A.O. Hero, “Dynamic stochastic blockmodels for time-evolving social networks,” *IEEE Journal on Selected Topics in Signal Processing*, vol. 8(4), pp.552-562, 2014.

[5] M. Frei and H.R. Künsch, “Mixture ensemble Kalman filters,” *Computational Statistics & Data Analysis*, vol. 58, pp. 127-138, 2013.

[6] A. Giménez, I. Khoury, J. Andrés-Ferrer, and J. Alfons, “Handwriting word recognition using windowed Bernoulli HMMs,” *Pattern Recognition Letters*, vol. 35, pp. 149-156, 2014.

[7] 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.

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

[9] A. Salazar and L. Vergara, *Independent Component Analysis (ICA): Algorithms, Applications and Ambiguities*. USA: Nova Science Publishers, 2018.

[10] 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.

[11] 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.

[12] A. Salazar, On Statistical Pattern Recognition in Independent Component Analysis Mixture Modelling. Springer-Verlag, Berlin, Heidelberg, 2013.

[13] G. Safont, A. Salazar, L. Vergara, E. Gomez, V. Villanueva, "Probabilistic Distance for Mixtures of Independent Component Analyzers," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 4, pp. 1161-1173, 2018.

[14] A. Salazar, L. Vergara, and R. Miralles, "On including sequential dependence in ICA mixture models," *Signal Processing*, vol. 90, pp. 2314-2318, 2010.

[15] G. Safont, A. Salazar, L. Vergara, and A. Rodriguez, "New Applications of Sequential ICA Mixture Models Compared with Dynamic Bayesian Networks for EEG Signal Processing," in *Fifth International Conference on Computational Intelligence, Communication Systems and Networks*, Madrid, Spain, 2013.

[16] O. Ibe, Markov Processes for Stochastic Modeling. Elsevier, London, UK, 2013.

[17] L. Xie and Z.-Q. Liu, “A coupled HMM approach to video-realistic speech animation,” *Pattern Recognition*, vol. 40, no. 8, pp. 2325-2340, 2007.

[18] P. Kumar, H. Gauba, P.P. Roy, and D.P. Dogra, “Coupled HMM-based multi-sensor data fusion for sign language recognition,” *Pattern Recognition Letters*, vol. 86, pp. 1-8, 2017.

[19] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9(8), pp. 1735-1780, 1997.

[20] R. Zhang, W. Yang, Z. Peng, P. Wei, X. Wang, L. Lin, “Progressively diffused networks for semantic visual parsing,” *Pattern Recognition*, vol. 90, pp. 78-86, 2019.

[21] M. Quintana et al., “Spanish multicenter normative studies (neuronorma project): norms for the abbreviated barcelona test,” *Archives of Clinical Neuropsychology*, vol. 26(2), pp. 144-157, 2011.

[22] M. Benedet and M. Alejandre, Test de Aprendizaje Verbal España-Complutense, Spain: TEA Ediciones, 1998.

[23] K.L. Gwet, Handbook of Inter-Rater Reliability. Advanced Analytics LLC, Gaithersburg, MD, USA, 2014.

[24] T.P. Jung and T.W. Lee, "Applications of Independent Component Analysis to Electroencephalography," in Statistical and Process Models for Cognitive Neuroscience and Aging. USA: Psychology Press, 2012.

[25] 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(2), pp. 391-403, 2011.

[26] P. Spurek, J. Tabor, P. Rola, and M. Ociepka, “ICA based on asymmetry,” *Pattern Recognition*, vol. 67, no. 1, pp. 230-244, 2017.

[27] A. Hyvärinen, "Statistical models of natural images and cortical visual representation," *Topics in Cognitive Science*, vol. 2(2), pp. 251-264, 2010.

[28] G. Safont, A. Salazar, and L. Vergara, "Multiclass Alpha Integration of Scores from Multiple Classifiers," *Neural Computation*, vol. 31(4), pp. 806-825, 2019.

[29] Z. Ghahramani and M.I. Jordan, “Factorial hidden Markov models,” in Advances in Neural Information Processing Systems, pp. 472-478, 1996.

[30] A. Hayashi, K. Iwata, and N. Suematsu, “Marginalized Viterbi algorithm for hierarchical hidden Markov models,” *Pattern Recognition*, vol. 46, no. 12, pp. 3452-3459, 2013.

[31] L.E. Baum, T. Petrie, G. Soules, and N. Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164-171, 1970.

[32] A.J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260-269, 1967.

[33] E. M. Thomas, A. Temko, W. P. Marnane, G. B. Boylan and G. Lightbody, "Discriminative and Generative Classification Techniques Applied to Automated Neonatal Seizure Detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 17(2), pp. 297-304, 2013.

[34] K.P. Burnjam and D.R. Anderson, Model Selection and Inference: A Practical Information-Theoretic Approach. Springer, NY, 2013.

[35] M. Kim, “Mixtures of Conditional Random Fields for Improved Structured Output Prediction," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28(5), pp. 1233-1240, 2017.

[36] H. Li, Q. Zhang, J. Deng, and Z.B. Xu, "A Preference-Based Multiobjective Evolutionary Approach for Sparse Optimization," *IEEE Transactions on Neural Networks and Learning Systems*, in press, doi: 10.1109/TNNLS.2017.2677973.

[37] C. Neuper and W. Klimesch, *Event-related dynamics of brain oscillations*. Amsterdam, NL: Elsevier, 2006.

[38] A. Megías, M.J. Gutiérrez-Cobo, R. Gómez-Leal, R. Cabello, P. Fernández-Berrocal, "Performance on emotional tasks engaging cognitive control depends on emotional intelligence abilities: An ERP study," *Scientific Reports*, vol. 7, no. 1, Article no. 16446, 2017.

[39] S. Rossi, E. Visani, F. Panzica, D. Sattin, A. Bersano, A. Nigri, S. Ferraro, E. Parati, M. Leonardi, and S. Franceschetti, "Sleep patterns associated with the severity of impairment in a large cohort of patients with chronic disorders of consciousness," *Clinical Neurophysiology*, vol. 129, no. 3, pp. 687-693, 2018.

[40] Y. Yau, Y. Zeighami, T.E. Baker, K. Larcher, U. Vainik, M. Dadar, V.S. Fonor, P. Hagmann, A. Griffa, B. Mišić, D.L. Collins, and A. Dagher, "Network connectivity determines cortical thinning in early Parkinson's disease progression," *Nature Communications*, vol. 9, no. 1, Article no. 2416, 2018.

[41] J.M. Antelis, L. Montesano, A. Ramos, and N. Birbaumer, "Decoding Upper Limb Movement Attempt from EEG Measurements of the Contralesional Motor Cortex in Chronic Stroke Patients," *IEEE Transactions on Biomedic Engineering*, available online, doi: 10.1109/TBME.2016.2541084, 2016.

[42] I. Daly, S.J. Nasuto, K. Warwick, “Brain computer interface control via functional connectivity dynamics,” *Pattern Recognitiion,* vol. 45, no. 6, pp. 2123-2136, 2012.

[43] S.L. Wendt, J.A.E. Christensen, J.Kempfner, H.L. Leonthin, P.Jennum and H.B.D. Sorensen, "Validation of a novel automatic sleep spindle detector with high performance during sleep in middle aged subjects," in *34th Annual International Conference of the IEEE EMBS*, pp. 4250-4253, San Diego (CA), 2012.

[44] S. Motamedi-Fakhr, M. Moshrefi-Torbati, M. Hill, C.M. Hill, and P.R. White, "Signal processing techniques applied to human sleep EEG signals-A review," *Biomedical Signal Processing and Control,* vol. 10, pp. 21-33, 2014.

[45] S. Sanei, J.A. Chambers, EEG Signal Processing. John Wiley & Sons, 2013.

[46] M.J. Dietz, K.J. Friston, J.B. Mattingley, A. Roepstorff, and M.I. Garrido, “Effective connectivity reveals right-hemisphere dominance in audiospatial perception: implications for models of spatial neglect,” *The Journal of Neuroscience*, vol. 34(14), pp. 5003-5011, 2014.

[47] Delorme, J. Palmer, J. Onton, R. Oostenveld, and S. Makeig, “Independent EEG sources are dipolar,” *PLoS ONE*, vol. 7(2), doi: 10.1371/journal.pone.0030135, 2012.

[48] R.T. Schirrmeiester, J.T. Springenberg, L.D.J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, “Deep learning with convolutional neural networks for EEG decoding and visualization,” *Human Brain Mapping*, vol. 38, pp. 5391-9420, 2017.

[49] J. Onton, A. Delorme, and S.Makeig, "Frontal midline EEG dynamics during working memory," *NeuroImage*, vol. 27, pp. 341-356, 2005.

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

[51] R. San-Segundo, R. Cordoba, J. Ferreiros, L.F. D'Haro-Enríquez, "Frequency features and GMM-UBM approach for gait-based person identification using smartphone inertial signals," *Pattern Recognition Letters*, vol. 73, pp. 60-67, 2017.