IDIAP RESEARCH REPORT

EEG Classification using Generative Independent Component Analysis

Silvia Chiappa and David Barber

IDIAP–RR 04-77

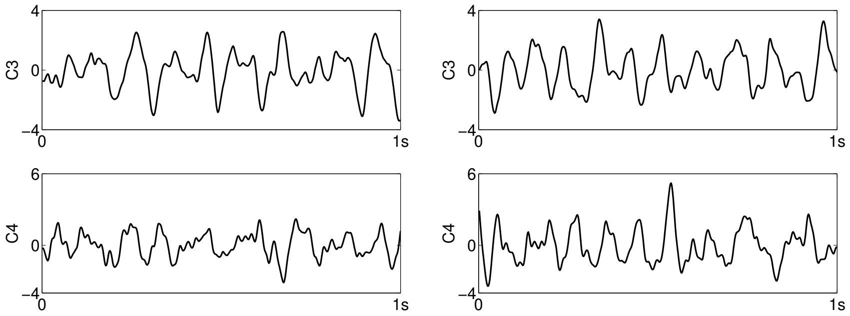
IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 04-77

EEG Classification using Generative Independent Component Analysis

Silvia Chiappa and David Barber

Abstract. We present an application of Independent Component Analysis (ICA) to the discrimination of mental tasks for EEG-based Brain Computer Interface systems. ICA is most commonly used with EEG for artifact identification with little work on the use of ICA for direct discrimination of different types of EEG signals. By viewing ICA as a generative model, we can use Bayes' rule to form a classifier. We fit spatial filters and source distribution parameters simultaneously and investigate whether these are sufficiently informative to produce good results when compared to more traditional methods based on using temporal features as inputs to off-the-shelf classifiers. Experiments suggest that state-of-the-art results may indeed be found without explicitly using temporal features. We extend the method to using a mixture of ICA models, consistent with the assumption that subjects may have more than one approach to thinking about a specific mental task.



(a) (b)

Figure 1: One second of EEG signal (in the band 6-26 Hz) recorded from electrodes C3 and C4 while a subject is performing (a) imagined left movement and (b) imagined right movement.

1 Introduction

EEG-based Brain Computer Interface (BCI) systems allow a person to control devices (such as a cursor on a screen) by using the electrical activity of the brain, recorded by electrodes placed over the scalp (see [23] for a general introduction on BCI research). In the case of systems based on endogenous brain activity, the user concentrates on different mental tasks (e.g. imagination of hand movement) which are associated with different device commands. The main envisaged use of EEG in this context is for persons with severe physical paralysis. An initial training phase (ideally as short as possible to avoid user fatigue) is required in order to calibrate the mental states realised by the user with a desired command. After this phase, ideally the subject will be able to reliably use the system actively for executing simple commands. However, the mental strategy taken varies widely across and also within subjects. BCI systems may therefore need to heavily adapt – possibly with instantaneous feedback – to the user. EEG is popular in this context since the system is portable and also has a fine temporal resolution (on the millisecond scale), enabling relatively rapid estimates of the subject's mental state [18]. Tasks are usually selected so that different brain areas become active while performing each specific task. A prominent characterization of activity is the attenuation of rhythmic components. For example, motor cortical areas which are not engaged in producing motor outputs often generate an EEG signal with rhythms in the α band (8-13 Hz) and, to a lesser extent, in the β band (18-26 Hz), called μ and β rhythms respectively. If a person moves his hand, the opposite-hemisphere cortical area becomes active and the rhythms diminish in that area. A similar effect occurs when a person *imagines* the movement, but no physical movement takes place [20].

Whilst EEG is demonstrably capable of containing meaningful information about the brain state, nevertheless, some important difficulties exist: the signals are relatively weak (in the range of 5-100 microvolts) and easily masked by noise such as mains-electrical interference; artifacts such as eye-movements and blinks, swallowing and other subject movements; inaccuracy of electrode placement; DC level (drift in the base activity of an electrode which is not correlated with the mental state and is an artifact of the instrumentation). In addition, other difficulties not specific to EEG arise, such as inconsistencies in the mental state the subject uses when asked to perform a particular task. These issues make the correspondence between electrode activity and mental state difficult to achieve reliably. In Fig. 1 we plot one second of typical EEG signal recorded from a subject performing (a) left and (b) right imaginary movements at two electrodes commonly used for discriminating these two mental tasks. The signal has been band-pass filtered between 6-26 Hz. No clear difference between the tasks is visually apparent and automatic procedures are required to perform task classification. For our machine learning approaches 17 electrodes were used to form the automatic classifiers. Standard approaches extract the frequency content of the signal in the α and β bands, which is then processed by a classifier. In many cases, a spatial filter is also applied to the data in order to extract more informative features. Popular approaches are based on Common Spatial Pattern algorithms [2, 22]. Another common approach is Independent Component Analysis (ICA) which transforms the raw signals into statistically independent sources. The temporal features of the spatially preprocessed data are then used as inputs to a standard classifier.

The central aim of this paper is to use directly a generative ICA model of EEG signals as a classifier. This is in sharp contrast to more traditional approaches which commonly view ICA-type methods only as a preprocessing step, with the exception of [19], where the authors introduce a combination of Hidden Markov Models and ICA as a generative model of the EEG data to detect switching between baseline activity and imaginary movement. Here we further investigate the use of a generative ICA model for EEG classification. However, we use a simplified model with no temporal dependence between the hidden sources, since we are here interested in whether or not the spatial information is a reliable indicator of the task, without the need to explicitly search for the presence of task-dependent temporal features. Two different datasets will be considered for analysis, classifying EEG signals based on word or movement tasks, as detailed in Section 3. Our approach will be to fit, for each person, an generative ICA model to each separate task, and then use Bayes' rule to form a classifier. The training criterion will be to maximise the class conditional likelihood. This will be compared with the more standard technique of using a Support Vector Machine (SVM) [6] trained with power spectral density features. We will compare two temporal feature types, one computed from raw data and the other from data preprocessed by a spatial filter.

2 Generative Independent Component Analysis (gICA)

From a basic understanding of the physics of the setup, linear ICA [10] seems an appropriate model of EEG signals and has been extensively applied to related tasks, such as the identification of artifacts and the analysis of the underlying brain sources [7, 14, 21]. Under the linear ICA assumption, signals v_t^j recorded at time $t = 1, \ldots, T$ at scalp electrodes $j = 1, \ldots, V$ are formed from a linear and instantaneous superposition of electromagnetic activity h_t^i in the cortex, generated by independent brain processes $i = 1, \ldots, H$, that is:

$$v_t = W h_t + \eta_t \,.$$

Here the mixing matrix W mimics the mixing and attenuation of the source signals. The term η_t potentially models additive measurement noise. For reasons of computational tractability[1], we consider here only the limit of zero noise. The empirical observations v_t are made zero-mean by a preprocessing step, which obviates the need for a constant output bias, and allows us to assume that h_t also has zero mean. Hence we can define $p(v_t|h_t) = \delta(v_t - W h_t)$, where $\delta(\cdot)$ is the Dirac Delta function. It is also convenient to consider square W, so that $V = H$. Our aim is to fit a model of the above form to each class of task c. In order to do this, we will describe each class specific model as a joint probability distribution, and use maximum likelihood as the training criterion. Whilst this is a hidden variable model ($h_{1:T_c}$ are hidden), thanks to the δ function, we can easily integrate out the hidden variables to form the likelihood of the visible variable $p(v_{1:T_c})$ directly [16], in contrast to the usual application of the EM algorithm in hidden variable models [17]. Given the above assumptions, the density of the observed and hidden variables for data from class c is

$$p(v_{1:T_c}, h_{1:T_c}|c) = \prod_{t=1}^{T_c} p(v_t|h_t, c) \prod_{i=1}^{H} p(h_t^i|c) = \prod_{t=1}^{T_c} \delta(v_t - W_c h_t) \prod_{i=1}^{H} p(h_t^i|c) \,. \tag{1}$$

Here $p(h_t^i|c)$ is the prior distribution of the activity of source i, and is assumed to be stationary. This forms a generative model of the output data v_t since one can first sample a value of the hidden vector

[1] Non zero noise may be dealt with at the expense of approximate inference [11].

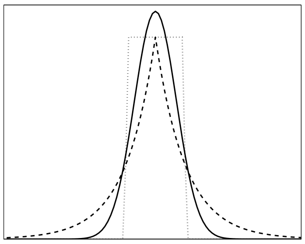


Figure 2: Generalized exponential distribution for $\alpha = 2$ (solid line), $\alpha = 1$ (dashed line) and $\alpha = 100$ (dotted line), which correspond to Gaussian, Laplacian and approximately uniform distributions respectively.

h_t, and then generate a visible vector using $v_t = W_c h_t$. By integrating (1) over the hidden variables h_t we obtain:

$$p(v_{1:T_c}|c) = \prod_{t=1}^{T_c} \int_{h_t} \delta(v_t - W_c h_t) \prod_{i=1}^{H} p(h_t^i|c) = |\det W_c|^{-T_c} \prod_{t=1}^{T_c} \prod_{i=1}^{H} p(h_t^i|c)\,, \tag{2}$$

where $h_t = W_c^{-1} v_t$.

There is an important difference between standard applications of ICA and the use of a generative ICA model for classification. In a standard usage of ICA, the sole aim is to estimate the mixing matrix W_c from the data. In that case, it is not necessary to model accurately the source distribution $p(h^i|c)$ [12]. Indeed, the statistical consistency of estimating W_c can be guaranteed using only two types of fixed prior distributions: one for modelling sub-Gaussian and another for modelling super-Gaussian h^i. However, the aim of our work is to perform classification, for which an appropriate model for the source distribution of each component h^i is fundamental.

As in [15, 19], we use the generalized exponential family which encompasses many types of symmetric and unimodal distributions[2]:

$$p(h^i|c) = \frac{f(\alpha_{ic})}{\sigma_{ic}} \exp\Big(-g(\alpha_{ic})\Big|\frac{h^i}{\sigma_{ic}}\Big|^{\alpha_{ic}}\Big)\,,$$

where

$$f(\alpha_{ic}) = \frac{\alpha_{ic}\Gamma(3/\alpha_{ic})^{1/2}}{2\Gamma(1/\alpha_{ic})^{3/2}}\,, \quad g(\alpha_{ic}) = \Big(\frac{\Gamma(3/\alpha_{ic})}{\Gamma(1/\alpha_{ic})}\Big)^{\alpha_{ic}/2}$$

and $\Gamma(\cdot)$ is the Gamma function. Although unimodality appears quite a restrictive assumption, our experience on the tasks we consider is that it is not inconsistent with the nature of the underlying sources, as revealed by a histogram analysis of $h_t = W_c^{-1} v_t$. The parameter σ_{ic} is the standard deviation[3], while α_{ic} determines the sharpness of the distribution as shown in Fig. 2. In the unconstrained case, where a separate model is fitted to data from each class independently, we aim to maximise the class-conditional log-likelihood

$$L(c) = \log p(v_{1:T_c}|c)\,.$$

In the case where parameters are tied across the different models, for example if the mixing matrix is kept constant over the different models ($W_c \equiv W$), the objective becomes instead $\sum_c L(c)$. Following the work in [19], we set to zero the derivatives of $L(c)$ with respect to σ_{ic}, obtaining the following

[2]Importantly, this is able to model both super and sub Gaussian distributions, which are required to isolate the independent components.

[3]Due to the indeterminacy of the variance of h_t^i (h_t^i can be multiplied by a scaling term a as long as the i^{th} column of W_c is multiplied by $1/a$), σ_{ic} could be set to one in the general model described above. However this cannot be done in the constrained version $W_c \equiv W$ considered in the experiments.

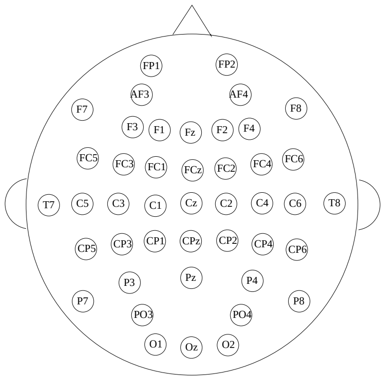


Figure 3: Electrode placement.

closed-form solution:

$$\sigma_{ic} = \left(\frac{g(\alpha_{ic})\alpha_{ic}}{T_c} \sum_{t=1}^{T_c} |h_t^i|^{\alpha_{ic}} \right)^{1/\alpha_{ic}} .$$

After substituting this optimal value of σ_{ic} into $L(c)$, the derivatives with respect to the parameters α_{ic} and W_c^{-1} are used in the scaled conjugate gradient method described in [1]. These are:

$$\frac{\partial L(c)}{\partial \alpha_{ic}} = \frac{T_c}{\alpha_{ic}} + \frac{T_c}{\alpha_{ic}^2} \frac{\Gamma'(1/\alpha_{ic})}{\Gamma(1/\alpha_{ic})} + \frac{T_c}{\alpha_{ic}^2} \log \left(\frac{\alpha_{ic} \sum_{t=1}^{T_c} |h_t^i|^{\alpha_{ic}}}{T_c} \right) - \frac{T_c \sum_{t=1}^{T_c} |h_t^i|^{\alpha_{ic}} \log |h_t^i|}{\alpha_{ic} \sum_{t=1}^{T_c} |h_t^i|^{\alpha_{ic}}}$$

$$\frac{\partial L(c)}{\partial W_c^{-1}} = T_c (W_c^\dagger - \sum_{t=1}^{T_c} b_t v_t^\dagger) , \quad \text{with} \quad b_t^i = \frac{\text{sign}(h_t^i) |h_t^i|^{\alpha_{ic}-1}}{\sum_{t=1}^{T_c} |h_t^i|^{\alpha_{ic}}} ,$$

where the prime symbol $'$ indicates differentiation and the $\dagger$ symbol indicates the transpose operator. After training, a novel test sequence $v_{1:T}^*$ is classified using Bayes' rule $p(c|v_{1:T}^*) \propto p(v_{1:T}^*|c)$, assuming $p(c)$ is uniform.

3 gICA versus SVM and ICA-SVM

3.1 Dataset I

This dataset concerns classification of the following three mental tasks[4]:

1. imagination of self-paced left hand movements,
2. imagination of self-paced right hand movements,
3. mental generation of words starting with a letter chosen spontaneously by the subject at the beginning of the task.

EEG potentials were recorded with the Biosemi ActiveTwo system [8], using the following electrodes located at standard positions of the 10-20 International System [13]: FP1, FP2, AF3, AF4, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T7, C3, Cz, C4, T8, CP5, CP1, CP2, CP6, CP6, P7, P3, Pz, P4, P8, PO3, PO4, O1, Oz and O2 (see Fig. 3). The raw potentials were re-referenced to the common average reference in which the overall mean is removed from each channel. The signals were recorded at a sample rate of 512 Hz. Subsequently, the band 6-26 Hz was selected with a 2nd order Butterworth filter. This preprocessing filter allow us to focus on μ and β rhythms. Experimentally,

[4]Available from www.idiap.ch/∼chiappa.

	Day 1					Day 2								
	Subjects A B C					Subjects A B				Subject C				
Training	1-2-3		4-5			1-2		3-4		1-2-3		4-5		
Validation	4	5	2-3	1-2	1-3	3	4	1	2	4	5	2-3	1-2	1-3
Testing	5	4	1	3	2	4	3	2	1	5	4	1	3	2

Table 1: Dataset I covers two days of data: 5 recording sessions on Day 1 for all subjects; for Day 2, Subjects A and B have 4 sessions and Subject C 5 sessions. The table describes how we split these sessions into training, validation and test sessions for the within-the-same-day experiments.

we also found that removing frequencies outside the band 6-26 Hz robustified the performance. Out of the 32 electrodes, only the following 17 electrodes were considered for the analysis: F3, Fz, F4, FC5, FC1, FC2, FC6, C3, Cz, C4, CP5, CP1, CP2, CP6, P3, PZ, P4 (see Fig. 3). This electrode selection was done on the basis of prior knowledge and a preliminary performance analysis. The data was acquired in an unshielded room from two healthy subjects without any previous experience with BCI systems. During an initial day the subjects familiarised themselves with the system, aiming to produce consistent mental states for each task. This data was not used for the training or analysis of the system. In the following two days several sessions were recorded for analysis, each session lasting around 4 minutes followed by an interval of around 5 to 10 minutes. Throughout all the training and test sessions, no feedback was provided to the subjects, neither in terms of the consistency of the mental states produced, nor results from automatic classification of the EEG signals. During each recording session, around every 20 seconds an operator verbally instructed the subject to continually perform one of the three mental tasks described above.

In a practical scenario, it is envisaged that a user will have an initial intense training period after which, ideally, very little retraining or re-calibration of the system should be required. The performance of BCI systems needs to be robust to potential changes in the manner that a subject performs a mental task from session to session, and indeed from day to day. Methods which are highly sensitive to such variations are unsuitable for a practical BCI system. We therefore performed two sets of experiments. In the first case, training, validation and testing were performed on data recorded within the same day, but using separate sessions. The detailed train, validation and test setting is given in Table 1. In the second set of experiments, we used the first day to train and validate the models, with test performance being evaluated on the second day alone and vice-versa. In particular, the first three sessions of one day were used for training and the last session(s) for validation.
Classification of the three mental tasks was performed using a window of one second of signal. That is, from each session we extracted around 210 samples of 512 frames, obtaining the following number of test examples: 1055, 1036 and 1040 for Day 1; 850, 836 and 1040 for Day 2 (subjects A, B and C respectively).

The non-temporal gICA model described in Section 2 was compared with two temporal feature approaches: the SVM and ICA-SVM. The purpose of these experiments is to consider whether or not using gICA can provide state-of-the-art performance compared to more standard methods based on using temporal features. Also of interest is whether or not standard ICA preprocessing would improve the performance of temporal feature classifiers.

gICA For gICA, no temporal features need to be extracted and the signal $v_{1:T}$ (downsampled to 64 samples per second) is used, as described in Section 2. Since we assume that the scalp signal is generated by a linear mixing of sources in the cortex, provided the data is acquired under the same conditions, it would seem reasonable to further assume that the mixing is the same for all classes ($W_c \equiv W$), and this constrained version was therefore also considered. The number of iterations for training the gICA parameters was determined using a validation set[5].

[5]The maximization of the log-likelihood (2) is a non-convex problem, thus the choice of the initial parameters may

Subject A	gICA W_c	gICA W	SVM	ICA-SVM
Train Day 1, Test Day 1	33.8±6.5%	34.7±5.8%	35.8±5.2%	34.7±5.5%
Train Day 2, Test Day 1	34.2±5.3%	36.1±5.0%	33.3±5.1%	32.8±5.6%
Train Day 2, Test Day 2	24.7±7.5%	26.8±7.1%	24.5±5.9%	25.1±6.3%
Train Day 1, Test Day 2	23.6±4.7%	24.6±5.0%	22.7±4.5%	24.0±2.4%
Subject B	**gICA W_c**	**gICA W**	**SVM**	**ICA-SVM**
Train Day 1, Test Day 1	31.4±7.1%	34.9±7.4%	38.4±5.2%	32.9±6.1%
Train Day 2, Test Day 1	45.6±5.1%	49.1±3.7%	42.1±4.7%	36.6±7.2%
Train Day 2, Test Day 2	32.5±4.4%	35.1±5.1%	36.7±3.0%	28.9±2.3%
Train Day 1, Test Day 2	31.4±2.3%	35.7±3.3%	39.3±4.3%	40.5±1.6%
Subject C	**gICA W_c**	**gICA W**	**SVM**	**ICA-SVM**
Train Day 1, Test Day 1	50.5±2.8%	49.4±4.2%	45.5±3.1%	49.0±3.4%
Train Day 2, Test Day 1	52.7±3.6%	55.7±3.3%	48.1±4.7%	52.5±3.8%
Train Day 2, Test Day 2	43.1±2.6%	45.0±4.2%	44.3±4.4%	44.8±3.5%
Train Day 1, Test Day 2	50.2±2.5%	55.3±4.2%	48.7±3.5%	54.9±2.9%

Table 2: Mean and standard deviation of the test errors in classifying three mental tasks using gICA with a separate W_c for each class (gICA W_c), gICA with a matrix W common to all classes (gICA W), SVM trained on PSD features (SVM) and SVM trained on PSD features computed from FASTICA transformed data (ICA-SVM). Random guessing corresponds to an average error of 66.7%.

SVM For the SVM method, we first need to find the temporal features which will subsequently be used as input to the classifier. Several power spectral density representations were considered. The best performance was obtained using Welch's periodogram method in which each pattern was divided into half-second length windows with an overlap of 1/4 of second, from which the average of the power spectral density (PSD) over all windows was computed. This gave a total of 186 feature values (11 for each electrode) as input for the classifier. Each class was trained against the others, and the kernel width (from 50 to 20000) and the parameter C (from 10 to 200) were found using the validation set.

ICA-SVM The data is first transformed by using the FASTICA algorithm [9] with the hyperbolic tangent nonlinearity and an initial W matrix equal to the identity, then processed as in the SVM approach above.

3.1.1 Results

A comparison of the performance of the spatial gICA against the more traditional methods using temporal features is shown in Table 2[6]. The setup of how exactly how each training and test sessions were used is given in Table 1. Together with the mean, we give the standard deviation of the error on the test sessions, which indicates the variability of performance obtained in different sessions. For

be important. We analyzed two cases in which the W_c matrix was initialized to the identity or to the matrix found by FASTICA [9] using the hyperbolic tangent (randomly initialized), while the exponents of the generalized exponential distribution α were set to 1.5. In both cases we obtained similar performance. We then decided to initialize W_c to the identity matrix in all subsequent experiments.

[6] A related version of this dataset also appeared in the third BCI competition [5]. However, there the task was based on a simpler same-day training and test situation (also with only a single test session), a larger classification window (1.44s) and 8 electrodes. The best results were found using a distance based classifier and an SVM with a Gaussian kernel, giving 31.3% and 31.5% error respectively. Whilst these results cannot be compared directly to the results in Table 2, they motivate the use of the Gaussian SVM in our comparative experiments.

gICA, using a different mixing matrix W_c for each mental task generally improves performance. Thus, in the following, we consider only gICA W_c for the comparison with the other standard approaches.

For Subject A, for which the best overall results are found, all three models give substantially the same performance, without loss when training and testing on different days.

For Subject B, for training and testing on the same day, gICA W_c and ICA-SVM perform similarly, and better than the SVM. However, when training on Day 2 and testing on Day 1, the performance of all models degenerates but more heavily for gICA W_c. ICA-SVM still gives some advantage over SVM. This situation is reversed when training on Day 1 and testing on Day 2.

For Subject C, the general performance of the methods is poor. Bearing this in mind, the SVM performs slightly better on average than gICA W_c and ICA-SVM when training and testing on the same day, whereas the two ICA models perform similarly. For training and testing on different days, on average, gICA slightly outperforms the ICA-SVM method, with the best results being given by the plain SVM method. A possible reason for this is that, in this subject, finding reliably the independent components is a challenging task with convergence difficulties often expressed by FASTICA, and the performance of the classifier may be hindered by this numerical instability.

In summary:

1. Training and testing on different days may significantly degrade performance. This indicates that some subjects may be either fundamentally inconsistent in their mental strategies, or the recording situation is not consistent. This more realistic scenario is to be compared with relatively optimistic results from more standard same-day training and testing benchmarks [5].

2. ICA preprocessing generally improves classification performance. However, in poorly performing subjects, the convergence of FASTICA was problematic, indicating that the ICA components were not reliably estimated, and thereby degrading performance.

3. gICA and ICA-SVM have similar overall performance. This indeed suggests that, for this dataset, state-of-the-art performance can be achieved using gICA, compared with temporal feature based approaches.

3.2 Dataset II

The second dataset analyzed in this work was provided for the BCI competition 2003 [3, 4]. This dataset differs from the previous one in that here the movements are real and not imagined, the assumption being that similar brain activity occurs when the corresponding movement is imagined only. The subject had to perform one of two tasks: depressing a keyboard key with a left or right finger.

EEG was recorded from one healthy subject during 3 sessions lasting 6 minutes each. Sessions were recorded during the same day at intervals of some minutes. The key depression occurred in a self-chosen order and timing. For the competition, 416 epochs of 500ms EEG were provided, each ending 130ms before an actual key press, at a sampling rate of 1000 and 100 Hz. The epochs were randomly shuffled and split into a training-validation set and a test set consisting of 316 and 100 epochs respectively. EEG was recorded from 28 electrodes covering the primary sensory motor area: F3, F1, Fz, F2, F4, FC5, FC3, FC1, FCz FC2, FC4, FC6, C5, C3, C1, Cz, C2, C4, C6, CP5, CP3, CP1, CPz, CP2, CP4, CP6, O1 and O2 (see Fig. 3).

The synchronous protocol used to record this data makes possible to consider, in addition to μ and β rhythms, another important EEG feature related to movement planning, called the Bereitschaftspotential (BP). BP is a slowly decreasing cortical potential which develops 1-1.5 seconds prior to a self-paced movement. The BP shows larger amplitude contralateral to the moving finger. The difference in the spatial distribution of BP is thus an important indicator of left or right finger movement. Indeed, the particular temporal shape of the BP may also be specific to the task and be a useful feature to aid classification. In order to include such a feature in the ICA or gICA approach, it is likely that a non-symmetric prior (or a non symmetric FASTICA approach) would need to be considered. To keep this paper relatively focused, we will apply only the symmetric gICA (and FASTICA) models to a preprocessed form of this dataset in which we filter to consider only μ-β bands, thereby removing any large scale shape effects such as the BP[7]. For the other methods not solely based on ICA, we retained possible BP features for a point of comparison to see if the use of BP features indeed is critical for reasonable performance on this database. The following methods were considered:

μ-β-gICA The μ-β filtered data is used as input to the generative ICA model described in Section 2.

BP-SVM This method focuses on the use of the BP as the features for a classifier. Here we preprocessed raw data in the 'BP band' (350 dimensional feature vector, 25 for each of the 14 electrodes). A Gaussian kernel was used and its width learned (in the range 10-5000), together with the strength of the margin constraint C (in the range 10-200), on the basis of the validation set.

μ-β-SVM This method focuses on the μ-β band, which precludes therefore any use of a BP for classification. The data was first filtered in the μ-β band as described above. Then the power spectral density was computed (168 dimensional feature vector).

BP-μ-β-SVM Here the combination of BP features and μ-β spectral features were used as input to an SVM classifier.

μ-β-ICA-SVM Here the μ-β filtered data is further preprocessed using FASTICA to form features to the SVM classifier.

BP-μ-β-ICA-SVM Here the combination of BP features with μ-β-ICA features forms the input to the SVM classifier.

3.2.1 Results

The comparison between these models is given in Table 3, in which we present the mean test error and standard deviation obtained by using 5-fold cross-validation[8]. Given the low number of test samples, it is difficult to present decisive conclusions. However, by comparing μ-β-SVM and μ-β-ICA-SVM, we note that using an ICA decomposition on μ-β filtered data improves performance. For this dataset, gICA-type models obtain superior performance to methods in which ICA is used as preprocessing. Finally, and perhaps most interestingly, the performance of gICA on μ-β is comparable with the results obtained by *combining* μ-β and BP features (BP-μ-β-ICA-SVM). The results from the gICA method are comparable to the best results previously reported for this dataset[9].

[7] We analyzed 100 Hz sampled data. The raw potentials were re-referenced to the common average reference. Then, the following 14 electrodes were selected: C5, C3, C1, Cz, C2, C4, C6, CP5, CP3, CP1, CPz, CP2, CP4 and CP6. For analyzing μ and β rhythms, each epoch was zero-mean and filtered in the band 10-32 Hz with a 2nd order Butterworth (zero-phase forward and reverse) digital filter. For BP, each epoch was low-pass filtered at 7 Hz using the same filtering setting, then the first 25 frames of each epoch were disregarded. This pre-processing was based on a preliminary analysis taking into consideration the best performance obtained in the BCI competition 2003 on this dataset [22].

[8] For each of the methods, we split the training data into 5 sets and performed cross-validation for hyperparameters by training on 4 sets and validating on the fifth. The resulting model was then evaluated on the separate test set. This procedure was repeated for the other four combinations of choosing 4 training and 1 validation set from the 5 sets. The mean and standard deviation of the 5 resulting models (for each method) are then presented.

[9] The winner of the BCI competition 2003 applied a spatial subspace decomposition filter and Fisher discriminant analysis to extract three types of features derived from BP and μ-β rhythms, and used a linear perceptron for classification. The final accuracy on the test was 16.0% [22].

μ-β-gICA W	μ-β-gICA W_c	BP-SVM	μ-β-SVM
16.0±1.2%	17.0±2.3%	21.6±1.5%	25.4±3.1%
BP-μ-β-SVM	μ-β-ICA-SVM	BP-μ-β-ICA-SVM	
18.8±0.8%	22.2±2.3%	16.2±0.8%	

Table 3: Mean and standard deviation of the the test errors in classifying two finger movement tasks. Random guessing corresponds to an error of 50%.

4 Mixture of Generative ICA

Although the performance of gICA is reasonable, if used in any BCI system, it would still achieve far from perfect performance. Whilst the reason for this may simply be inherently noisy data, another possibility is that the subject's reaction when asked to think about a particular mental task drifts significantly from one session and/or day to another. It is also natural to assume that a subject has more than one way to think about a particular mental task. The idea of using a mixture model is to test the hypothesis that the data may be naturally split into regimes, within which a single model may accurately model the data, although this single model is not able to model accurately all the data. This motivates the following model for a single sequence of observations

$$p(v_{1:T_c}|c) = \sum_{m=1}^{M_c} p(v_{1:T_c}|m,c)p(m|c)\,,$$

where m describes the mixture component. The number of mixture components M_c will typically be rather small, being less than 5. We will then fit a separate mixture model to data for each class c. To ease the notation a little, from here we drop the class dependency. Training this model by maximising the likelihood directly is cumbersome. A more elegant approach is afforded by the EM algorithm [17], which enables us to perform maximum likelihood in the context of latent or hidden variables, in this case being played by m. EM is an iterative procedure which, at each iteration, computes the set of parameters (in our case $\{\sigma_{im},\alpha_{im},W_m$ and $p(m)\}$) which maximises the so-called expectation of the complete data log-likelihood, computed using the parameters from the previous iteration. In the mixture case we have a set of sequences $v^s_{1:T}$, $s = 1,\ldots,S$ each of the same length T. The expected complete data log-likelihood is given by

$$\begin{aligned} L &= \left\langle \log \prod_{s=1}^{S} p(v^s_{1:T}|m)p(m) \right\rangle_{p(m|v^s_{1:T})} \\ &= \sum_{s=1}^{S} \left\langle \sum_{t=1}^{T} \log|\det W_m^{-1}|p(W_m^{-1}v^s_t) + \log p(m) \right\rangle_{p(m|v^s_{1:T})}, \qquad (3) \end{aligned}$$

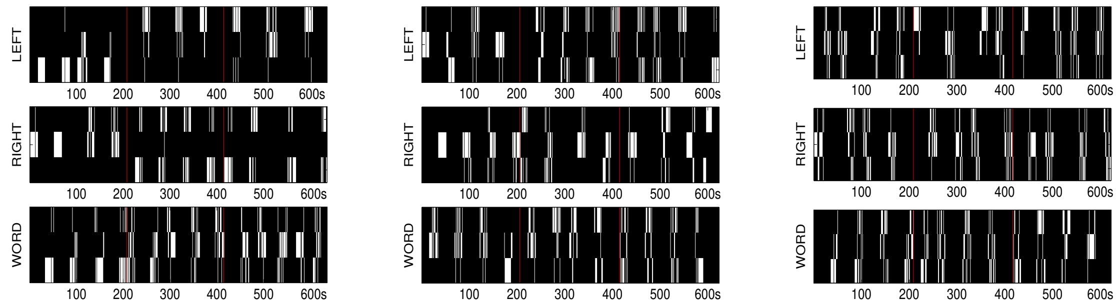
where S indicates the number of sequences and $\langle\cdot\rangle$ indicates the expectation operator. Here v^s_t is the vector of observations at time t from sequence s. At each iteration of the EM algorithm, the prior is updated as

$$p(m) = \frac{1}{S}\sum_{s=1}^{S} p(m|v^s_{1:T})\,,$$

where

$$p(m|v^s_{1:T}) = \frac{p(v^s_{1:T}|m)p(m)}{\sum_{m'=1}^{M} p(v^s_{1:T}|m')p(m')}\,.$$

The other parameters are then updated analogously to the single component case by computing the derivatives of Equation (3).



Subject A Subject B Subject C

Figure 4: We show here results of fitting a separate mixture model with three components to each of the three tasks for the first three sessions of Day 1. Time (in seconds) goes from left to right. At any time, only one of the three classes (corresponding to the verbal instruction to the subject), and only one of the three hidden states for that class (the one with the highest posterior probability), is highlighted in white. The plot shows how the subjects change in their strategy for realising a particular mental task with time. The vertical lines indicate the boundaries of the training sessions, which correspond to a gap of 5-10 minutes.

4.1 gICA versus Mixture of gICA

4.1.1 Dataset I

We first fitted a mixture of three gICA models to the first three sessions of Day 1. The aim here is that this may enable us to visualise how each subject switches between its mental strategies, and therefore perhaps to form an idea of how reliably each subject is performing. These results are presented in Fig. 4, where switching for each subject between the three different mixture components is shown. Interestingly, we see that for Subjects A and B and all three tasks, only a single component tends to be used during the first session, suggesting a high degree of consistency in the way that the mental tasks were realised. For Subject C, a lesser degree of reliability is present. This situation changes so that, in the latter two sessions, a much more rapid switching occurs (indeed this switching happens much more quickly than the time prescribed for a mental task). This suggests that the consistency with which subjects perform the mental tasks deteriorates with time, highlighting the need to potentially account for such drift in approach.

To see whether or not this results in an improved classification, we trained the mixture of gICA model, as described above, on the dataset. Table 4 compares the performance between gICA and the mixtures of gICA models using a separate W_c matrix for each class. The number of mixture components (ranging from 2 to 5) was chosen from the validation set. The W_c was initialized adding a small amount of noise to W_c found using one mixture. Whilst the mixture of ICA model seems to be reasonably well motivated, disappointingly, only a minor improvement with respect to the single mixture case is found on Subjects A and B. It is not clear why the performance improvement is so modest. This may be due to the fact that whilst drift is indeed an issue and better modelled by this approach, the model does not capture the online nature of adaptation that may occur in practice. That is, a stationary mixture model may be inadequate for capturing the dynamic nature of changes in user mental strategies.

Subject A	gICA W_c	MgICA W_c
Train Day 1, Test Day 1	33.8±6.5%	31.1±4.9%
Train Day 2, Test Day 1	34.2±5.3%	33.6±5.0%
Train Day 2, Test Day 2	24.7±7.5%	22.3±6.4%
Train Day 1, Test Day 2	23.6±4.7%	22.4±3.0%
Subject B	gICA W_c	MgICA W_c
Train Day 1, Test Day 1	31.4±7.1%	30.6±3.8%
Train Day 2, Test Day 1	45.6±5.1%	40.0±10.0%
Train Day 2, Test Day 2	32.5±4.4%	29.1±3.0 %
Train Day 1, Test Day 2	31.4±2.3%	29.5±6.0 %
Subject C	gICA W_c	MgICA W_c
Train Day 1, Test Day 1	50.5±2.8%	52.2±4.8%
Train Day 2, Test Day 1	52.7±3.6%	52.2±2.7%
Train Day 2, Test Day 2	43.1±2.6%	44.6±3.2%
Train Day 1, Test Day 2	50.2±2.5%	51.6±1.6%

Table 4: Mean and standard deviation of the test errors in classifying three mental tasks using gICA with a separate W_c for each class (gICA W_c) and a mixture of gICA with a separate W_c for each class (MgICA W_c).

4.1.2 Dataset II

The result of using a mixture model with a separate W_c for each class is $19.4 \pm 2.6\%$. Compared with the results presented from the single gICA and other methods in Table 3, this result is disappointing, being a little (though not significantly) worse than the single gICA method. Here, the number of mixture components (from 2 to 5) is chosen on the basis of the validation set and this should, in principle, avoid overfitting. However, the validation error for a single component is often a little better than for a number of mixture components greater than 1, suggesting indeed that the model is overfitting slightly.

5 Conclusions

In this work we have presented an analysis on the use of a spatial generative Independent Component Analysis (gICA) model for the discrimination of mental tasks for EEG-based BCI systems. We have compared gICA against other standard approaches, where temporal information from a window of data (power spectral density) is extracted and then processed using an SVM classifier. Our results suggest that using gICA alone is powerful enough to produce good performance for the datasets considered. Furthermore, using ICA as a preprocessing step for power spectral density SVM classifiers also tends to improve the performance, giving roughly the same performance as gICA. An important point is that performance generally degrades when one trains a method on one day and tests on another, although for some subjects this is less apparent. This more realistic scenario is a more severe test of BCI methods and, in our view, merits further consideration. For this reason, we investigated whether or not a mixture model, which may cope with potentially severe changes in mental strategy, may improve performance. Indeed, the use of mixture models appears to be well-founded since, based on the training data alone, switching between mixture components tends to increase with time. However the resulting performance improvements for classification were rather modest (or even slightly worse), suggesting that the model is overfitting slightly. Indeed, the model does not deal well with the potentially dynamic nature of change. An online version of training may be a reasonable way to avoid this difficulty, by which some form of continual recalibration based on feedback is provided.

Acknowledgement

This work is supported by the Swiss National Science Foundation NCCR IM2 and by the European IST Programme FET Project FP6-003758. This paper only reflects the authors views and funding agencies are not liable for any use that may be made of the information contained herein.

References

[1] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[2] G. Blanchard and B. Blankertz. BCI competition 2003 - Data set IIa: spatial patterns of self-controlled brain rhythm modulations. *IEEE Transactions on Biomedical Engineering*, 51:1062–1066, 2004.

[3] B. Blankertz, G. Curio, and K.-R. Müller. Classifying single trial EEG: Towards brain computer interfacing. In *Advances in Neural Information Processing Systems*, volume 14, pages 157–164, 2002.

[4] B. Blankertz, K.-R. Müller, G. Curio, T. M. Vaughan, G. Schalk, J. R. Wolpaw, A. Schölgl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schröder, and N. Birbaumer. The BCI competition 2003: progress and perspectives in detection and discrimination of EEG single trials. *IEEE Transactions on Biomedical Engineering*, 51:1044–1051, 2004.

[5] BCI competition III. ida.first.fraunhofer.de/projects/bci/competition_iii, 2005. Dataset V.

[6] N. Cristianini and J. S. Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[7] A. Delorme and S. Makeig. EEG changes accompanying learned regulation of 12-Hz EEG activity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11:133–137, 2003.

[8] http://www.biosemi.com.

[9] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10:626–634, 1999.

[10] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, Inc., 2001.

[11] Pedro Højen-Sørensen, Lars Kai Hansen, and Ole Winther. Mean field implementation of Bayesian ICA. In *3rd International Conference on Independent Component Analysis and Blind Signal Separation*, pages 439–444, 2001.

[12] J.-F. Cardoso. On the stability of source separation algorithms. In *Workshop on Neural Networks for Signal Processing*, pages 13–22, 1998.

[13] H. H. Jasper. Ten-twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology*, 10:371–373, 1958.

[14] T. P. Jung, C. Humphries, T. W. Lee, S. Makeig, M. J. McKeown, V. Iragui, and T. Sejnowski. Extendend ICA removes artifacts from electroencephalografic recordings. *Advances in Neural Information Processing Systems*, 10:894–900, 1998.

[15] T.-W Lee and M. S Lewicki. The generalized Gaussian mixture model using ICA. In *International Workshop on Independent Component Analysis*, pages 239–244, 2000.

[16] D. J. C. MacKay. Maximum likelihood and covariant algorithms for independent component analysis. Cambridge University, Cavendish Laboratory, 1999.

[17] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. John Wiley and Sons, 1997.

[18] P. L. Nunez. *Neocortical Dynamics and Human EEG Rhythms*. Oxford University Press, 1995.

[19] W. D. Penny, S. J. Roberts, and R. M. Everson. Hidden Markov independent components for biosignal analysis. In *International Conference on Advances In Medical Signal and Information Processing*, pages 244–250, 2000.

[20] G. Pfurtsheller and C. Neuper. *Movement and ERD/ERS*, pages 191–206. Kluwer Academic/Plenum Publishers, 2003.

[21] R. Vigário. Extraction of ocular artefacts from EEG using independent components analysis. *Electroencephalography and Clinical Neurophysiology*, 103:395–404, 1997.

[22] Y. Wang, Z. Zhang, Y. Li, X. Gao, S. Gao, and F. Yang. BCI competition 2003-Data set IV: An algorithm based on CSSD and FDA for classifying single-trial EEG. *IEEE Transactions on Biomedical Engineering*, 51:1081–1086, 2004.

[23] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113:767–791, 2002.