Generative Temporal ICA for Classification in Asynchronous BCI Systems

Silvia Chiappa and David Barber

IDIAP Research Institute,
Switzerland

Abstract— **In this paper we investigate the use of a temporal extension of Independent Component Analysis (ICA) for the discrimination of three mental tasks for asynchronous 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. In a recent work we have shown that, by viewing ICA as a generative model, we can use Bayes' rule to form a classifier obtaining state-of-the-art results when compared to more traditional methods based on using temporal features as inputs to off-the-shelf classifiers. However, in that model no assumption on the temporal nature of the independent components was made. In this work we model the hidden components with an autoregressive process in order to investigate whether temporal information can bring any advantage in terms of discrimination of spontaneous mental tasks.**

I. Introduction

EEG-based Brain Computer Interface (BCI) systems allow a person to control devices by using observed electrical activity v_t^j, at time t, recorded by electrodes placed over the scalp at locations $j = 1, \ldots, V$. In the case of systems based on spontaneous brain activity, the user concentrates on different mental tasks (e.g. imagination of hand movement) which are associated with different device commands. Tasks are normally selected so that task-dependent areas in the brain become active. The most prominent characterization of activity is the attenuation of rhythmic components, mostly in the α band. Standard approaches extract the frequency content of the signal, which is then processed by a static classifier (see [12] for a general introduction on BCI research).

Signals recorded at scalp electrodes are commonly considered as a linear and instantaneous superposition of unobserved or hidden electromagnetic activity h_t^i generated by independent brain processes, $i = 1, \ldots, H$. For these reasons Independent Component Analysis (ICA) [6] seems an appropriate model of EEG signals and has been extensively applied to related tasks, such as the identification of artifacts ([7], [11]) and the analysis of the underlying brain sources.

More specifically related to BCI research, several studies have addressed the issue of whether an ICA decomposition can enhance differences in the mental tasks such as to improve the performance of brain-actuated systems. Most of these studies use static versions of ICA either as a form of preprocessing, or to aid analysis of the signal. In contrast to our approach below, they do not use ICA itself to directly form a classifier. In [8], the authors analyze a visual attention task and show that ICA finds μ-components which show a spectral reactivity to motor events stronger than the one measured from scalp channels. They suggest that ICA can be used for optimizing brain-actuated control. In [3] ICA is used for analyzing EEG data recorded from subjects which attempt to regulate power at 12 Hz over the left-right central scalp. Other studies use ICA as a denoising technique or as a feature extractor for improving the performance of a separate classifier. For example, in [4] ICA is used to remove ocular artefacts, while [5] extracts task-related independent components prior the application of several classifiers. In contrast to these approaches, in [10] the authors introduce a combination of Hidden Markov Models and Independent Component Analysis as a generative model of the EEG data and give a demonstration of how this model can be applied directly to the detection of when switching occurs between the two mental conditions of baseline activity and imaginary movement.

Following a similar approach, in a recent work [2] we have used directly a simple static ICA generative model of EEG signals as a classifier for the recognition of three mental tasks. We have shown that a performance similar to standard approaches based on using temporal features as inputs to off-the-shelf classifiers can be obtained. It is still an open question whether we can do better by using a more complex model of the data, since in [2] the temporal nature of the independent components was not taken into account. Temporal modeling of the hidden components, for example with autoregressive models [9], has shown to improve separation in the case of other types of recordings.

In this paper we further investigate the use of ICA for classification by modeling each hidden component with an autoregressive process. Our interest is to asses performance in experiments which are close to the real use of a BCI system. Rather than using a synchronous protocol, in our system the subject performs repetitive movements and word generation in a self-paced manner, without being synchronized to an external cue.

Our approach is to fit, for each person, an ICA generative model to each separate task, and then use Bayes' rule to form directly a classifier. This model will be compared with its static special case, where no temporal information is taken into account, and with two standard techniques for the recognition of mental tasks: the Multilayer Perceptron (MLP)

This work was supported by the Swiss NSF through the NCCR IM2 and by the PASCAL Network of Excellence, IST-20002-506778, funded in part by the Swiss OFES.

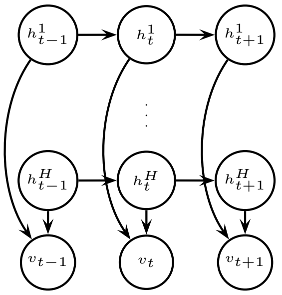

Fig. 1. Graphical representation of an ICA model with temporal dependence between the hidden variables (of order $p = 1$).

and Support Vector Machine (SVM) [1], trained with power spectral density features.

II. Generative Temporal Independent Component Analysis

Generative Independent Component Analysis is a probabilistic model in which a vector of observations v_t is assumed to be generated by statistically independent (hidden) random variables h_t via an instantaneous linear transformation:

$$v_t = W h_t + \epsilon_t \,,$$

where ϵ_t is noise. For reasons of tractability, in our model (and others in the literature) ϵ_t will be assumed to be zero throughout, and W will be assumed to be a square matrix.

Like in Contextual ICA [9] and HMMICA [10], we assume temporal dependence between the hidden variables h_t by modeling the i^{th} hidden brain process h^i_t with a linear autoregressive model of order p:

$$h^i_t = \sum_{k=1}^{p} a^i_k h^i_{t-k} + \eta^i_t = \hat{h}^i_t + \eta^i_t \,,$$

where η^i_t is the noise term. Graphically, the Bayesian network which corresponds to this model is shown in Fig. **??**.
Our aim will be to fit a model of the above form to each class of task c. In order to do this, we will describe the model as a joint probability distribution, and use maximum likelihood as the training criterion.
Given the above assumptions, we can factorize the density of the observed and hidden variables as follow[1]:

$$p(v_{1:T}, h_{1:T}|c) = \prod_{t=1}^{T} p(v_t|h_t, c) \prod_{i=1}^{H} p(h^i_t|h^i_{t-1:t-p}, c) \,. \quad (1)$$

Using $p(v_t|h_t) = \delta(v_t - W h_t)$ we can easily integrate (1) over the hidden variables h_t to form the likelihood of the observed sequence $v_{1:T}$:

$$p(v_{1:T}|c) = |\det W_c|^{-T} \prod_{t=1}^{T} \prod_{i=1}^{H} p(h^i_t|h^i_{t-1:t-p}, c) \,, \quad (2)$$

[1]This is a slight abuse for reasons of notational simplicity. The model is only defined for $t > p$. This is true for all subsequent dependent formulae.

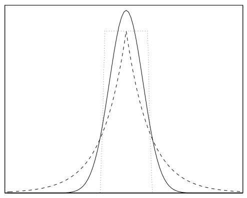

Fig. 2. Generalized exponential distribution for $\alpha = 2$ (solid line), $\alpha = 1$ (dashed line) and $\alpha = 100$ (dotted line), which correspond to Gaussian, Laplacian and approximately uniform distributions respectively.

where $h_t = W_c^{-1} v_t$.
We will model $p(h^i_t|h^i_{t-1:t-p}, c)$ with the generalized exponential distribution:

$$p(h^i_t|h^i_{t-1:t-p}, c) = \frac{f(\alpha^{ic})}{\sigma^{ic}} \exp\Big(- g(\alpha^{ic}) \Big| \frac{h^i_t - \hat{h}^i_t}{\sigma^{ic}} \Big|^{\alpha^{ic}} \Big) \,,$$

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. The generalized exponential family encompasses many types of symmetric and unimodal distributions. The parameter σ is the standard deviation[2], while α determines the sharpness of the distribution, as shown in Fig. 2.
The logarithm of the likelihood (2) is summed over all training sequences belonging to each class and then maximized by using the scaled conjugate gradient method described in [1]. This requires computing the derivatives with respect to all the parameters, that is, the mixing matrix W_c, the autoregressive coefficients a^i_k, and the parameters of the exponential distribution σ^{ic} and α^{ic} (see APPENDIX).
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.

III. Experimental Setup

EEG potentials were recorded with the Biosemi ActiveTwo system (http://www.biosemi.com), using 32 electrodes located at standard positions of the 10-20 International System, at a sample rate of 512 Hz. The raw potentials were re-referenced to the Common Average Reference in which the overall mean is removed from each channel. Subsequently, the band 6-16 Hz was selected with a Butterworth filter. This preprocessing filter is a simple way to remove strong drift terms in the signals (the so-called DC level) and the 50 Hz noise, which are artifacts of instrumentation and do not correspond to brain activity. Experimentally, we also found that removing frequencies outside the band 6-16 Hz robustified the performance. Only the following 19 electrodes were considered for the analysis: F3, FC1, FC5, T7, C3, CP1, CP5, P3, Pz, P4, CP6, Cp2, C4, T8, FC6, FC2, F4, Fz and Cz.

The data were acquired in an unshielded room from two healthy subjects without any previous experience with BCI

[2]Due to the indeterminacy of variance of the h^i (h^i can be multiplied by a scaling term a as long as the corresponding column of W_c is multiplied by $1/a$), σ could be set to one in the general model described above. However this cannot be done in the constrained version $W_c = W$ considered in the experiments (see Sec. III).

systems. During an initial day the subjects learned how to perform the mental tasks. In the following two days, 10 recordings, each lasting around 4 minutes, were acquired for the analysis. During each recording session, every 20 seconds an operator instructed the subject to perform one of three different mental tasks. The tasks were: (1) imagination of self-paced left, (2) right hand movement and (3) mental generation of words starting with a given letter.

IV. Results

The time series obtained from each recording session was split into segments of signal lasting one second. ICA was compared with two standard approaches, in which for each segment the power spectral density was extracted and then processed using an MLP and a SVM. The best performance was obtained using the following Welch's periodogram method: each pattern was divided into a quarter of second long windows with an overlap of 1/8 of second. Then the overall average was computed.

The first three sessions of each day were used for training the models while the other two sessions where used alternatively for validation and testing.

A softmax, one hidden layer MLP was trained using cross-entropy, with the validation set used to choose the number of iterations, the number of $\tanh$ hidden units (ranging from 1 to 100) and the learning rate of the gradient ascent method.

In the SVM, each class was trained against the others, and the standard deviation for the Gaussian SVM found using the validation set (ranging from 1 to 20000).

In the ICA model, for computational expediency only, the data were down-sampled from 512 to 64 samples per second. The validation set was used to choose the number of conjugate gradient iterations and the order p of the autoregressive model (from 1 to 8), even if we have observed that the appropriate order does not change for different sessions. Since we assume that the scalp signal is generated by linear mixing of sources in the cortex, provided the data are 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 is also considered.

A comparison of the performance is shown in Table I. Besides the results obtained with the Temporal ICA model (T. ICA), in which the independent components are modeled by an autoregressive process, we present the results obtained with a Static ICA model (S. ICA), which can be seen as a particular case in which the autoregressive order p is set to zero. Classification is measured on around 420 test examples. ICA consistently performs as well as the temporal feature approach using MLP and SVMs. However, by modeling the independent components with an autoregressive process we don't obtain improvements in discrimination with respect to the static case.

For Subject A, we used the third day's data to select the three hidden components whose distribution varied most across the three classes, using the ICA model with a matrix W common to all classes. In the Static ICA model, the three components were selected by looking at the distribution $p(h_t^i)$, while in the Temporal ICA model they were selected by looking at the conditional distribution $p(h_t^i|h_{t-1:t-p}^i)$ for the order p that gave the best performance in the test set. The projection of each component on the 19 scalp electrodes (i^{th} column of W) gives an indication of which part of the scalp received more activity from that component. The scalp projections and time courses (300 frames of the word task) of the selected hidden components are shown in Fig. 3. As we can see from the projections, there is a correspondence between the static components (s1, s2, s3) and temporal components (t1, t2, t3). The time courses are also very similar. In general we have found a high correspondence among almost all the 19 components of the Static and Temporal ICA model. The components for which a correspondence was not found don't show differences in the autoregressive coefficients and in the conditional distribution, thus are not relevant for discrimination. Finally note that the hidden components found by the Temporal ICA don't look smoother as we would expect.

TABLE I

Classifications errors for three mental tasks using Static ICA, Temporal ICA, MLP and SVM. W_c uses a separate matrix for each class, as opposed to a common matrix W.

	Subject A		Subject B	
	Day 2	Day 3	Day 2	Day 3
S. ICA W	40.0%	34.8%	28.5%	31.5%
T. ICA W	40.2%	36.7%	27.8%	30.8%
S. ICA W_c	37.1%	36.0%	25.6%	30.8%
T. ICA W_c	38.8%	36.2%	27.1%	28.2%
MLP	37.1%	38.1%	30.5%	34.2%
SVM	35.1%	38.1%	32.4%	36.6%

V. Conclusions

In this work we have presented a preliminary analysis on the use of a simple temporal Independent Component Analysis model for the discrimination of three mental tasks for asynchronous EEG-based BCI systems. Unlike standard static ICA, which assumes temporal independence of the hidden components, we have modeled each component with an autoregressive process. While this approach has been successfully applied to the separation of sources not well separable using static ICA, it does not seem to bring additional discriminant information when ICA is used as a generative model for direct classification. The reason may be that a simple linear model is not suitable for our EEG data, due to strong non-stationarity in the hidden dynamics. It may be more appropriate to use a switching model which can handle changes of regime in the EEG dynamics.

Acknowledgment

The authors would like to acknowledge Dr. S. Bengio and C. Dimitrakakis for useful discussions.

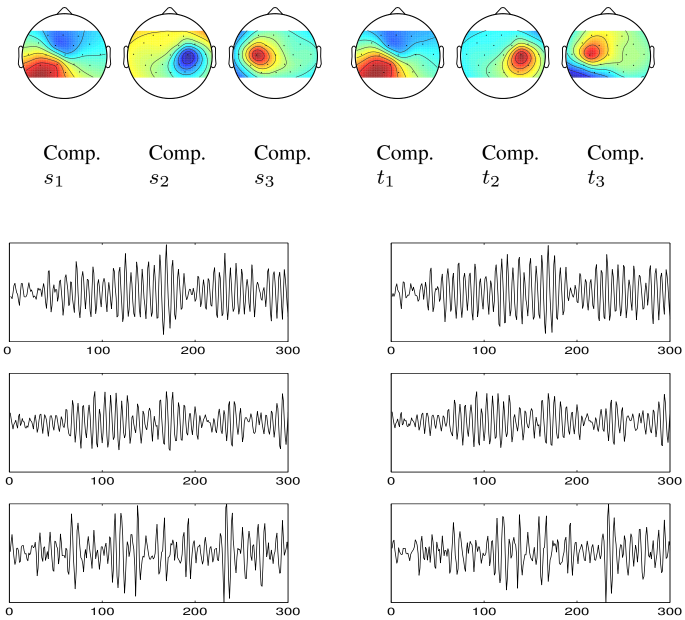

Fig. 3. Projection on the scalp of three hidden components for Subject A, Day 3 using Static ICA (Comp. s1, Comp. s2, Comp. s3) and Temporal ICA (Comp. t1, Comp. t2, Comp. t3) (From blue to red, negative to positive values). The topographic plots have been obtained by interpolating the values at the electrode (black dots) using the open source eeglab toolbox (http://www.sccn.ucsd.edu/eeglab). Below the projections, time courses (300 frames) of the corresponding hidden components. Due to the indeterminacy of variance of the hidden components, axes scale between different figures cannot be compared and has been removed. This also applies to the absolute scalp projection.

APPENDIX

Here we write the normalized log-likelihood of a set of

$$L(c) = \frac{1}{S_c(T-p)} \sum_{s=1}^{S_c} \log p(v^s_{p+1:T} | h^s_{1:p}, c)\,,$$

where s indicates the s^{th} training pattern of class c. We write $p(v^s_{p+1:T}|h^s_{1:p}, c)$, rather than the notational abuse $p(v^s_{1:T}|c)$ in the main text, since this takes care of the initial time steps which would otherwise be problematic. In the following, $h^s_t = W_c^{-1} v^s_t$, for $t = 1, \dots, T$. We want to maximize $L = \sum_c L(c)$. Dropping the pattern index s, the component index i and the class index c we have:

$$\frac{\partial L}{\partial \sigma} = -\frac{1}{\sigma} + \frac{g(\alpha)\alpha}{S(T-p)|\sigma|^{\alpha+1}} \sum_{s=1}^{S} \sum_{t=p+1}^{T} |h_t - \hat{h}_t|^{\alpha}\,,$$

that is the maximum likelihood solution is:

$$\sigma = \left(\frac{g(\alpha)\alpha}{S(T-p)} \sum_{s=1}^{S} \sum_{t=p+1}^{T} |h_t - \hat{h}_t|^{\alpha} \right)^{1/\alpha}.$$

Using this solution we obtain:

$$\frac{\partial L}{\partial \alpha} = \frac{1}{\alpha} + \frac{1}{\alpha^2} \frac{\Gamma(1/\alpha)'}{\Gamma(1/\alpha)} + \frac{1}{\alpha^2} \log \left(\frac{\alpha \sum_{s=1}^{S} \sum_{t=p+1}^{T} |h_t - \hat{h}_t|^{\alpha}}{S(T-p)} \right) - \frac{\sum_{s=1}^{S} \sum_{t=p+1}^{T} |h_t - \hat{h}_t|^{\alpha} \log |h_t - \hat{h}_t|}{\alpha \sum_{s=1}^{S} \sum_{t=p+1}^{T} |h_t - \hat{h}_t|^{\alpha}}.$$

Setting $A = W^{-1}$:

$$\frac{\partial L}{\partial A} = -\frac{1}{S(T-p)} \sum_{s=1}^{S} \sum_{t=p+1}^{T} b_t v_t' + \frac{1}{S(T-p)} \sum_{s=1}^{S} \sum_{t=p+1}^{T} \hat{B}_t + (A')^{-1}\,,$$

where b_t is a vector of elements

$$b^i_t = \frac{g(\alpha^i)}{(\sigma^i)^{\alpha^i}} \alpha^i \mathrm{sign}(h^i_t - \hat{h}^i_t) |h^i_t - \hat{h}^i_t|^{\alpha^i - 1}$$

and $\hat{B}_t$ is a matrix of rows

$$\hat{B}^i_t = \frac{g(\alpha^i)}{(\sigma^i)^{\alpha^i}} \alpha^i \mathrm{sign}(h^i_t - \hat{h}^i_t) |h^i_t - \hat{h}^i_t|^{\alpha^i - 1} \sum_{k=1}^{p} a^i_k v'_{t-k}\,.$$

Finally, the derivative with respect to the autoregressive coefficient a_k is:

$$\frac{\partial L}{\partial a_k} = \frac{g(\alpha)}{S(T-p)(\sigma)^{\alpha}} \alpha \sum_{s=1}^{S} \sum_{t=p+1}^{T} \mathrm{sign}(h_t - \hat{h}_t) |h_t - \hat{h}_t|^{\alpha - 1} h_{t-k}\,.$$

REFERENCES

[1] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford Univ. Press, 1995.
[2] S. Chiappa and D. Barber. Generative independent component analysis for EEG classification. Technical Report IDIAP-RR 77, IDIAP Research Institute, Martigny, Switzerland, 2004.
[3] A. Delorme and S. Makeig. EEG changes accompanying learned regulation of 12-Hz EEG activity. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 11:133–137, 2003.
[4] T. Hoya, G. Hori, H. Bakardjian, T. Nishimura, T. Suzuki, Y. Miyawaki, A. Funase, and J. Cao. Classification of single trial EEG signals by a combined principal + independent component analysis and probabilistic neural network approach. In *International Symposium on Independent Component Analysis and Blind Signal Separation*, pages 197–202, 2003.
[5] C.-I Hung, P.-L. Lee, Y.-T. Wu, H.-Y. Chen, L.-F Chen, T.-C. Yeh, and J.-C. Hsieh. Recognition of motor imagery electroencephalography using independent component analysis and machine classifiers. In *International Conference on Computer Graphics, Visualization and Computer Vision*, 2004.
[6] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, Inc., 2001.
[7] 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*, pages 894–900, 1998.
[8] S. Makeig, S. Enghoff, T.-P. Jung, and T. J. Senjnowski. A natural basis for efficient brain-actuated control. *IEEE Transactions on Rehabilitation Engineering*, 8:208–211, 2000.
[9] B. A. Pearlmutter and L. C. Parra. Maximum likelihood blind source separation: A context-sensitive generalization of ICA. In *Advances in Neural Information Processing Systems*, pages 613–619, 1997.
[10] 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.
[11] R. Vigário. Extraction of ocular artefacts from EEG using independent components analysis. *Electroencephalography and Clinical Neurophysiology*, 103:395–404, 1997.
[12] 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.