

Generative Independent Component Analysis for EEG Classification

Silvia Chiappa and David Barber *

IDIAP Research Institute,
Switzerland

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. This enables us also to investigate whether simple spatial information is sufficiently informative to produce state-of-the-art results when compared to more traditional methods based on using temporal features as inputs to off-the-shelf classifiers. Experiments conducted on two subjects suggest that knowing 'where' activity is happening alone gives encouraging results.

1 Introduction

EEG-based Brain Computer Interface (BCI) systems allow a person to control devices by using the electrical activity of the brain, recorded by electrodes placed over the scalp. 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 usually selected so that different brain areas become active while performing each one. In addition to 'where' activity is, 'what' the activity (or absence of activity) is may also be characteristic for a certain task. A prominent characterization of activity is the attenuation of rhythmic components, mostly in the α band (8-13 Hz). Standard approaches extract the frequency content of the signal, which is then processed by a static classifier (see [6] for a general introduction on BCI research). In this paper we try to answer the question whether the discrimination of mental tasks can be based essentially on spatial information alone.

Signals v_t^j recorded at time t at scalp electrodes $j = 1, \dots, V$ are commonly considered as a linear and instantaneous superposition of electromagnetic activity h_t^i in the cortex, generated by independent brain processes $i = 1, \dots, H$. For these reasons ICA 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.

The central aim of this paper is to use directly a simple generative ICA model of EEG signals as a classifier. This is in sharp contrast to more traditional

*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. The authors would like to thank S. Bengio for discussions.

approaches, which commonly view ICA-type methods only as a preprocessing step, with the exception of [5], where the authors introduce a combination of Hidden Markov Models and ICA 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. Using this paper as a basis, we further investigate the use of ICA for classification. However, we use a simplified model with no temporal dependence between the hidden sources h_t^i , since we are here interested critically 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. Our approach will be to fit, for each person, an ICA generative model to each separate task, and then use Bayes' rule to form directly a classifier. This will be compared with two more standard techniques: the Multilayer Perceptron (MLP) and Support Vector Machine (SVM) [1], trained with power spectral density features.

2 Generative Independent Component Analysis

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

$$v_t = Wh_t + \eta_t .$$

For reasons of computational tractability, we restrict ourselves to the limit of zero noise $\eta_t = 0$. Hence $p(v_t|h_t) = \delta(v_t - Wh_t)$, where $\delta(\cdot)$ is the delta function. It is also convenient to consider square W , so that $V = H$. Unlike [4, 5], we assume temporal independence between the hidden variables h_t .

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 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 follows:

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

Here $p(h_t^i|c)$ is the prior distribution of the activity of source i , and is assumed to be stationary. By integrating (1) over the hidden variables h_t we obtain:

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

where $h_t = W_c^{-1}v_t$. As is well known, it is not necessary to accurately model the source distribution $p(h_t^i|c)$ in order to correctly estimate W_c [2]. Indeed, statistical consistency of estimating W_c can be guaranteed using only two types of

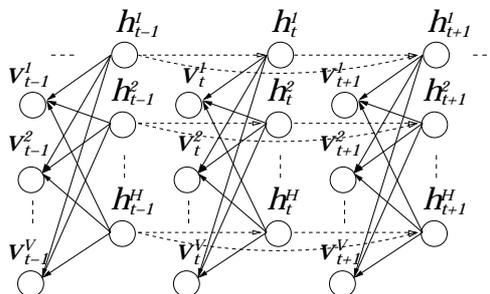


Fig. 1: Graphical representation of the Generative ICA model. The dashed lines indicate temporal dependence between the hidden variables, not considered in our model.

fixed prior distributions: one for modeling sub-Gaussian and another for modeling super-Gaussian h_t^i . However the aim of this work is to perform classification, for which an appropriate model for the source distribution is fundamental. As in [3, 5], we use the generalized exponential family which encompasses many types of symmetric and unimodal distributions¹:

$$p(h_t^i|c) = \frac{f(\alpha^{ic})}{\sigma^{ic}} \exp\left(-g(\alpha^{ic})\left|\frac{h_t^i}{\sigma^{ic}}\right|^{\alpha^{ic}}\right),$$

where

$$f(\alpha^{ic}) = \frac{\alpha^{ic}\Gamma(3/\alpha^{ic})^{1/2}}{2\Gamma(1/\alpha^{ic})^{3/2}}, \quad g(\alpha^{ic}) = \left(\frac{\Gamma(3/\alpha^{ic})}{\Gamma(1/\alpha^{ic})}\right)^{\alpha^{ic}/2}$$

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 the histogram analysis of $h_t = W_c^{-1}v_t$. The parameter σ is the standard deviation², while α determines the sharpness of the distribution³.

The logarithm of (2) is summed over all training patterns belonging to each class and maximized using the scaled conjugate gradient method described in [1]⁴.

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.

¹We zero mean the data, hence we can assume that the distribution is zero mean.

²Due 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 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. 3).

³ $\alpha < 2$, $\alpha = 2$, $\alpha > 2$ describe super-Gaussian, Gaussian or sub-Gaussian pdfs respectively.

⁴Let define $L = \sum_c \frac{1}{S_c T} \sum_{s=1}^{S_c} \log p(v_{1:T}^s|c)$, where s indicates the s^{th} training pattern of class c . In order to maximize L we compute the derivatives with respect to the parameters σ^{ic} , α^{ic} and W_c . Dropping the pattern index s , the component index i and the class index c we have:

$$\frac{\partial L}{\partial \sigma} = -\frac{1}{\sigma} + \frac{g(\alpha)\alpha \text{sign}(\sigma)}{ST|\sigma|^{\alpha+1}} \sum_{s=1}^S \sum_{t=1}^T |h_t|^\alpha, \quad \text{that is } |\sigma|^\alpha = \frac{g(\alpha)\alpha}{ST} \sum_{s=1}^S \sum_{t=1}^T |h_t|^\alpha.$$

3 Experiments

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 19 of the 32 electrodes, namely those covering the temporal-motor cortex were considered for the analysis (see Fig. 2).

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 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.

The time series obtained from each recording session was split into segments of signal lasting half (one) second. ICA was compared with two standard approaches, in which for each segment the power spectral density was extracted and then processed using a (softmax) MLP and a SVM [1]⁵. The first three sessions of each day were used for training the models while the other two sessions were used alternatively for validation and testing⁶. 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 ($W_c = W$) and this

Using this maximum-likelihood solution we obtain:

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

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

$$\frac{\partial L}{\partial A} = (A')^{-1} - \sum_{s=1}^S \sum_{t=1}^T b_t v_t', \quad \text{where } b_t^i = \frac{\text{sign}(h_t^i) |h_t^i|^{\alpha^i - 1}}{\sum_{s=1}^S \sum_{t=1}^T |h_t|^\alpha}.$$

⁵The best performance was obtained using the following Welch's periodogram method: each pattern was divided into a quarter of second length windows with an overlap of 1/8 of second. Then the average of the power spectral density over all windows was computed.

⁶A 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. In the SVM, each class was trained against the others, and the standard deviation (from 1 to 20000) for the Gaussian SVM was found using the validation set. In the ICA model, for computational expediency only, the data was down-sampled from 512 to 64 samples per second. The validation set was used to choose the number of iterations.

constrained version is also considered.

A comparison of the performance of our spatial ICA method against the more traditional methods using temporal features is shown in Table 1. ICA consistently performs as well as the temporal feature approach using MLP and SVM.

For each subject, we used one day's data to select the two hidden components h_t^i whose distribution varied most across the three classes, using the ICA model with a matrix W common to all classes. The projection of each component on the 19 channels (i^{th} column of W) gives an indication of which part of the scalp received more activity from that component. The distributions and scalp projections are shown in Fig. 2. Visually, the projections of components a_1 and b_1 are most similar. For these two components, the word task (dashed line) has the strongest activation (width of the distribution), followed by the left task (solid line) and the right task (dotted line). Gratifyingly, this suggests that for these two subjects a similar spatial pattern of activity occurs when they are asked to perform the tasks. To a lesser extent, visually components a_2 and b_2 are similar in their scalp projection, and again the order of class activation in the two components is the same (word task followed by right and left tasks).

	Subject A				Subject B			
	Day 2		Day 3		Day 2		Day 3	
	1/2 s	1 s	1/2 s	1 s	1/2 s	1 s	1/2 s	1 s
ICA W	42.1%	40.0%	39.3%	34.8%	33.5%	28.5%	35.8%	31.5%
ICA W_c	40.9%	37.1%	39.8%	36.0%	31.5%	25.6%	35.8%	30.8%
MLP	44.9%	37.1%	40.4%	38.1%	40.3%	30.5%	44.6%	34.2%
SVM	39.6%	35.1%	42.0%	38.1%	43.0%	32.4%	39.4%	36.6%

Table 1: Test errors in classifying three mental tasks using ICA with a matrix W common to all classes (ICA W), ICA with a separate W for each class (ICA W_c), MLP and SVM. The first, second column of each day indicates the error rate using half, one second of data respectively (around 840 (420) test examples).

4 Conclusions

In this work we have presented a preliminary analysis on the use of a purely spatial Independent Component Analysis model for the discrimination of mental tasks for EEG-based BCI systems. We have compared ICA with two other standard approaches, where temporal information from a window of data (power spectral density) is extracted and then processed using a static classifier. Our results suggest that spatial information alone is indeed powerful enough to produce state-of-the-art performance.

More sophisticated ICA approaches which take into account temporal information have been proposed in the literature. For example, in [4] the hidden components are modeled with an autoregressive process. It would be interesting to investigate whether this information can bring any advantage in terms of discrimination. Additionally, more complex source distributions may bring

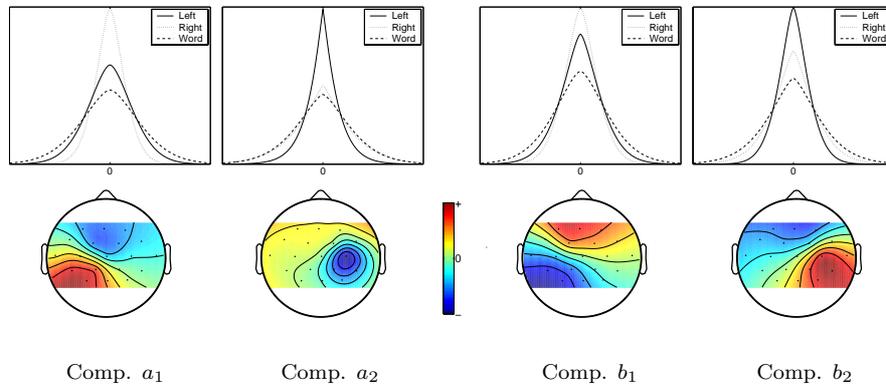


Fig. 2: Estimated pdfs and scalp projection of two hidden components for Subject A, Day 3 (Comp. a1, Comp. a2) and Subject B, Day 2 (Comp. b1, Comp. b2). The topographic plots have been obtained by interpolating the values at the electrode (black dots) using the eeglab toolbox (<http://www.sccn.ucsd.edu/eeglab>). Due to the indeterminacy of the hidden component variance, axes scale between figures cannot be compared and has been removed. This also applies to the scalp projection.

performance benefits. A key research issue is how to avoid using an initial filtering preprocessing step and make a consistent generative model of the raw data signal.

References

- [1] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford Univ. Press, 1995.
- [2] J.-F. Cardoso. On the stability of source separation algorithms. In *Workshop on Neural Networks for Signal Processing*, pages 13–22, 1998.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.