Blind Source Separation by Entropy Rate Minimization

Germán Gómez-Herrero, *Student Member, IEEE,* Kalle Rutanen, and Karen Egiazarian, *Senior Member, IEEE*

Abstract

An algorithm for the blind separation of mutually independent and/or temporally correlated sources is presented in this paper. The algorithm is closely related to the maximum likelihood approach based on entropy rate minimization but uses a simpler contrast function that can be accurately and efficiently estimated using nearest-neighbor distances. The advantages of the new algorithm are highlighted using simulations and real electroencephalographic data.

Index Terms

Blind Source Separation, Independent Component Analysis, Electroencephalography

I. INTRODUCTION

In this article we consider the simplest and most common Blind Source Separation (BSS) scenario:

$$\mathbf{x}(n) = \sum_{j=1}^{M} \mathbf{a}_j s_j(n) = \mathbf{A}\mathbf{s}(n) \tag{1}$$

where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_M]$ is an unknown $M \times M$ mixing matrix of full rank, $\mathbf{s}(n) = [s_1(n), \ldots, s_M(n)]^T$ are the sources and $\mathbf{x}(n) = [x_1(n), \ldots, x_M(n)]^T$ the observed mixtures. The index $n = 1, \ldots, N$ denotes discrete time instants. The goal of BSS is to estimate a separating matrix $\mathbf{B}$ such that the source signals can be approximately recovered up to a permutation and scaling indeterminacy, i.e. $\mathbf{BA} \approx \mathbf{P}\mathbf{\Lambda}$ where $\mathbf{P}$ and $\mathbf{\Lambda}$ are an arbitrary permutation matrix and an arbitrary diagonal matrix, respectively. A major

Germán Gómez-Herrero, Kalle Rutanen and Karen Egiazarian are with the Department of Signal Processing, Tampere University of Technology, Finland (e-mail: german.gomezherrero@tut.fi, kalle.rutanen@tut.fi, karen.egiazarian@tut.fi).

Manuscript received ??, ??; revised ??, ??.

application of BSS is the separation of brain sources from electroencephalography (EEG) and its magnetic counterpart, MEG [1].

At least three classes of source models have been proposed for which very efficient BSS algorithms exist: non-Gaussian [2], spectrally diverse [3], or non-stationary [4] independent sources. Algorithms unifying several of these source assumptions are especially promising for neuroscientific applications [5], [6].

In this article we introduce a novel BSS contrast which simultaneously exploits non-Gaussianity and temporal self-dependencies of the sources. Based on this contrast we develop a BSS algorithm which is robust to the presence of time-lagged cross-dependencies between sources. Such cross-dependencies are likely to occur in EEG/MEG applications due to (time-delayed) axonal propagation of information across distributed brain areas.

II. Definition of the BSS contrast

For sources that behave like i.i.d. non-Gaussian random variables, the linear and instantaneous BSS problem in Eq. 1 can be solved using Independent Component Analysis (ICA) [7]. If the observed mixtures have a covariance matrix $\mathbf{\Sigma_x} = E\left[\mathbf{x}\mathbf{x}^T\right]$, the ICA-based separating matrix is $\mathbf{B}_{opt} = \mathbf{R}_{opt}\mathbf{\Sigma_x}^{-1/2}$ where $\mathbf{R}_{opt}$ is the $M \times M$ unitary matrix that minimizes:

$$L_H(\mathbf{R}) = \sum_{i=1}^{M} H\left(\mathbf{r}_i \mathbf{\Sigma_x}^{-1/2}\mathbf{x}\right) = \sum_{i=1}^{M} H(\hat{s}_i) \tag{2}$$

where $\mathbf{R} = [\mathbf{r}_1, \ldots, \mathbf{r}_M]^T$ and H denotes Shannon entropy. But in most practical applications, and especially in the case of EEG, the sources are not i.i.d. and are better modeled as stochastic processes with (second and higher-order) temporal correlations. Let us consider sources that behave like mutually independent stationary Markov processes of order d so that their temporal structure is confined within the vector $\mathbf{s}_i^{(d)} = [s_i(n), s_i(n-1), \ldots, s_i(n-(d-1))]$. Note that, due to the stationarity of s_i, the joint distribution of $\mathbf{s}_i^{(d)}$ is invariant with respect to shifts in the time index. The amount of temporal structure in s_i (i.e. its temporal predictability) can be assessed by its *entropy rate* [8]:

$$H_r(s_i) = H\left(\mathbf{s}_i^{(d)}\right) - H\left(\mathbf{s}_i^{(d-1)}\right) \tag{3}$$

Indeed, in the case of Markovian sources of order d, a maximum likelihood estimate of the separating matrix is obtained by minimizing the entropy rate of the estimated sources [9]. However, estimating the entropy rate requires the combination of two different estimates of joint entropy, which increases the final estimation error. We take here a simpler approach to entropy rate minimization that consists on minimization of the following BSS contrast:

$$L(\mathbf{R}) = \sum_{i=1}^{M} H\left(\hat{\mathbf{s}}_i^{(d)}\right) \tag{4}$$

Intuitively, minimizing Eq. 4 will lead to source estimates which are maximally non-Gaussian (i.e. spatially independent) and that have maximum temporal dependencies. Eq. 4 is a valid BSS contrast because if $\mathbf{s}_i^{(d)}$ and $\mathbf{s}_j^{(d)}$ are mutually independent and at least one of them do not follow the Normal law, then the following inequality holds [8], [10]:

$$L\left(\alpha\mathbf{s}_i^{(d)} + \beta\mathbf{s}_j^{(d)}\right) \geq \min\left(L\left(\mathbf{s}_i^{(d)}\right), L\left(\mathbf{s}_j^{(d)}\right)\right)$$

with equality if and only if $\alpha = 0$ and $L\left(\mathbf{s}_j^{(d)}\right) < L\left(\mathbf{s}_i^{(d)}\right)$, $\beta = 0$ and $L\left(\mathbf{s}_i^{(d)}\right) < L\left(\mathbf{s}_j^{(d)}\right)$, or $\alpha\beta = 0$ and $L\left(\mathbf{s}_j^{(d)}\right) = L\left(\mathbf{s}_i^{(d)}\right)$. A generalization of this inequality to more than two sources can be found in [11].

If the sources are not perfectly independent, the global minimum of independence-based contrasts, like the one proposed here, might be spurious [10]. This is a crucial issue in the analysis of EEG data since brain sources are likely to exchange information through time-delayed axonal pathways. Thus, time-lagged cross-dependencies between the observed signals $x_1, \ldots, x_M$ should be minimized before using any independence-based contrast for BSS. Moreover, pervasive autocorrelations like those found in EEG time-series can also have a negative impact by effectively increasing the order of the Markov model that best fits the sources. A straightforward solution to these two problems is to use a vector autoregressive (VAR) model to filter out temporal correlations in the data [1]. Since the mixing matrix $\mathbf{A}$ in Eq. 1 commutes with the VAR filter, the filtered observations can be interpreted as a linear mixture of the filtered sources with the same mixing matrix. Our numerical experiments confirm that VAR filtering has generally positive effects on independence-based contrasts, especially in the separation of EEG sources. A detailed analysis of the benefits and drawbacks of VAR pre-processing will be presented elsewhere.

III. Estimation and optimization

Evaluation of the BSS contrast in Eq. 4 requires the estimation of multivariate Shannon entropies. We use an estimator based on k-nearest-neighbor distances [12]:

$$\hat{H}\left(\hat{\mathbf{s}}_i^{(d)}\right) = h_{k-1} - h_{N-1} + dE\left[\log 2\epsilon\right] \tag{5}$$

where $h_t = -\sum_{r=1}^{t} r^{-1}$, ϵ is the maximum norm distance from $\hat{\mathbf{s}}_i^{(d)}$ to its k-nearest neighbor and $E[\cdot]$ is the expectation operator, which can be approximated by the sample mean. In the numerical experiments below we always used $k = 20$ and $d = 4$. We preferred this estimator instead of fixed-bandwidth kernel-based approaches due to its higher sensitivity to finer (high-order) details of the distributions [13]. The optimum rotation minimizing the contrast in Eq. 4 is found using Jacobi rotations [7]. The resulting ENRICA (ENtropy Rate-based ICA) algorithm can be summarized in the following steps:

1) Whiten the data through the transformation $\mathbf{z} = \hat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1/2}\mathbf{x}$ where $\hat{\mathbf{\Sigma}}_{\mathbf{x}} = \frac{1}{N-1}\sum_{i=1}^{N}\mathbf{x}(n)\mathbf{x}^T(n)$.
2) Perform a temporal whitening by fitting a VAR model to $\mathbf{z}$ using ARfit [14] and computing the residuals of the model: $\mathbf{v}(n)\ \forall n = 1, \ldots, N$.
3) Using Jacobi rotations, find the unitary matrix $\hat{\mathbf{R}} = [\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2, \ldots, \hat{\mathbf{r}}_M]^T$ minimizing $\sum_{i=1}^{M} \hat{H}\left(\hat{\mathbf{s}}_i^d\right)$, where $\hat{s}_i(n) = \hat{\mathbf{r}}_i\mathbf{v}(n)$.
4) Estimate the separating matrix as $\hat{\mathbf{B}} = \hat{\mathbf{R}}\hat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1/2}$.

IV. Numerical experiments

In this section we assess the effectivity of ENRICA for separating (i) mutually independent sources with non-linear temporal dynamics, (ii) sources with time-lagged cross-dependencies and (iii) mutually independent EEG sources. In each of these scenarios we compared ENRICA with three groups of BSS algorithms[1]:

- Algorithms commonly used with EEG data: Infomax [15], EfICA [2] and WASOBI [3].
- Algorithms based on non-parametric estimates of entropy: RADICAL [16], MILCA [13] and NpICA [17]
- Algorithms that exploit simultaneously temporal structure and spatial independence: MCOMBI [6], ThinICA [5].

We tested the performance of all benchmark algorithms with and without preprocessing VAR filter in order to assess objectively the merits of ENRICA's contrast. In the first set of experiments we generated the hidden sources using three identical Lorenz oscillators $\mathbf{\Phi}_i : \left(\dot{X}_i(t), \dot{Y}_i(t), \dot{Z}_i(t)\right)$, $\forall i = 1, 2, 3$, described by the differential equations:

[1]We only considered algorithms for which public implementations were available.

$$
\begin{aligned}
\dot{X}_i(t) &= 10\left(Y_i(t) - X_i(t)\right) \\
\dot{Y}_i(t) &= 28X_i(t) - Y_i(t) - X_i(t)Z_i(t) \\
&\quad + \textstyle\sum_{j\neq i} K_{ij}\left(Y_j(t-\tau_{ji})\right)^2 \\
\dot{Z}_i(t) &= X_i(t)Y_i(t) - \tfrac{8}{3}Z_i(t)
\end{aligned}
$$

We integrated these equations with a time step of 0.3 and the sources were obtained from the Y components of the oscillators, i.e. $s_i(t) = Y_i(t)\ \forall i = 1,2,3$. Each source contained 3000 samples. We considered the case of uncoupled ($K_{ij} = 0\ \forall i,j$) and unidirectionally coupled oscillators ($K_{21} = 1$, $K_{32} = 1$, $K_{ij} = 0$ otherwise). In the latter case the coupling delays were $\tau_{21} = 10$ and $\tau_{32} = 15$. We generated 200 realizations of the mixtures by using different initial conditions for the Lorenz systems and random well-conditioned mixing matrices. Separation accuracy was assessed using the median interference-to-signal ratio (ISR) [18]:

$$
\text{ISR} = \text{median}_i\left(10\log\left(\sum_{k\neq i}\mathbf{G}_{ik}^2/\mathbf{G}_{ii}^2\right)\right) \tag{6}
$$

where $\mathbf{G}_{ik} = \left(\hat{\mathbf{B}}\mathbf{A}\right)_{ik}$. In general, ISR values above -10 dBs are probably unacceptable in most applications. The results in table I show that the optimal separation matrix corresponded to a robust global minimum of ENRICA's contrast function. VAR filtering was quite effective in removing spurious global minima in ICA-based contrasts but was not able to remove the numerous local minima, which explains the bad results of ICA algorithms based on local optimization (Infomax, EfICA and NpICA).

In order to assess the expected performance on real EEG data we tested ENRICA on mixtures of three time-series extracted from a real EEG dataset [19]. Mutual independence was achieved by selecting, from different electrodes, EEG epochs that did not overlap in time. This approach ensured that the time-courses of the sources mimic the dynamics of the underlying brain sources. However, the total lack of cross-dependencies between sources is unrealistic and, therefore, the numerical results of this experiment should be taken as positively biased estimates. From Fig. 1 it is obvious that BSS based on temporal structure clearly outperformed i.i.d.-based approaches, at least for realistic sample sizes. The poor convergence of Infomax raises concerns on common practices among the EEG research community. For instance, [20] recommends to use about $30M^2$ samples to estimate M sources but, in our experiments, Infomax needed at least 10 times more samples to produce reliable source estimates. Fig. 2 shows the 90th percentile of the median ISR for mixtures of more than three EEG sources. Even in high-dimensional problems, ENRICA consistently outperformed the benchmark algorithms. The major disadvantage of ENRICA with respect

TABLE I: Accuracy of the tested algorithms in the blind separation of three Lorenz oscillators. The numbers in boldface are median values (across 200 random surrogates of the Lorenz mixtures) of the median ISR. The subindices and superindices denote the 2.5% and 97.5% percentile values, respectively. The rows marked with the term "+VAR" indicate that VAR filtering was used to pre-process the observed mixtures.

BSS Algorithm		ISR (dB)	
		uncoupled	coupled
Infomax [15]		$\mathbf{-13}^{0}_{-32}$	$\mathbf{0}^{0}_{-1}$
	+VAR	$\mathbf{-4}^{1}_{-17}$	$\mathbf{-2}^{-1}_{-5}$
EfICA [2]		$\mathbf{-7}^{0}_{-32}$	$\mathbf{0}^{0}_{-1}$
	+VAR	$\mathbf{-2}^{0}_{-9}$	$\mathbf{-2}^{-1}_{-22}$
WASOBI [3]		$\mathbf{-2}^{1}_{-14}$	$\mathbf{-17}^{-14}_{-21}$
	+VAR	$\mathbf{-3}^{1}_{-18}$	$\mathbf{-12}^{-1}_{-26}$
RADICAL [16]		$\mathbf{-32}^{-26}_{-38}$	$\mathbf{-1}^{0}_{-28}$
	+VAR	$\mathbf{-28}^{-21}_{-37}$	$\mathbf{-21}^{-1}_{-33}$
MILCA [13]		$\mathbf{-32}^{-26}_{-40}$	$\mathbf{-8}^{0}_{-26}$
	+VAR	$\mathbf{-26}^{-20}_{-35}$	$\mathbf{-18}^{-6}_{-30}$
NpICA [17]		$\mathbf{-33}^{0}_{-46}$	$\mathbf{-1}^{0}_{-33}$
	+VAR	$\mathbf{-29}^{0}_{-36}$	$\mathbf{-4}^{-2}_{-33}$
MCOMBI [6]		$\mathbf{-6}^{0}_{-31}$	$\mathbf{-24}^{-1}_{-34}$
	+VAR	$\mathbf{-3}^{0}_{-13}$	$\mathbf{-2}^{-1}_{-28}$
ThinICA [5]		$\mathbf{-30}^{-24}_{-41}$	$\mathbf{-1}^{0}_{-19}$
	+VAR	$\mathbf{-28}^{-23}_{-38}$	$\mathbf{-18}^{-8}_{-28}$
ENRICA []		$\mathbf{-34}^{-29}_{-42}$	$\mathbf{-34}^{-27}_{-42}$
	+VAR	$\mathbf{-34}^{-29}_{-41}$	$\mathbf{-33}^{-26}_{-40}$

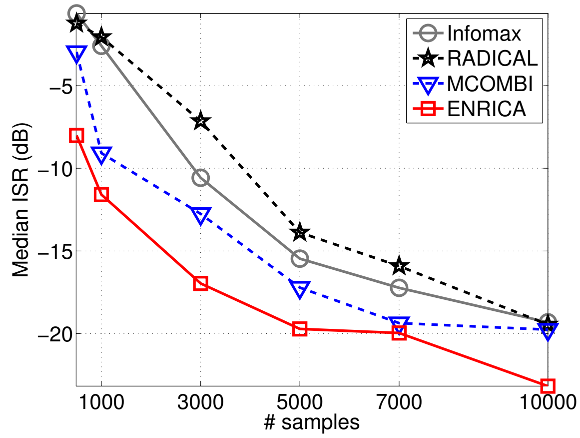

Fig. 1: 90th percentile of the median ISR (in dBs) across 200 random surrogates of the three EEG sources. For clarity we show only the results of the best performing algorithms. Full results of this experiment are available online [21].

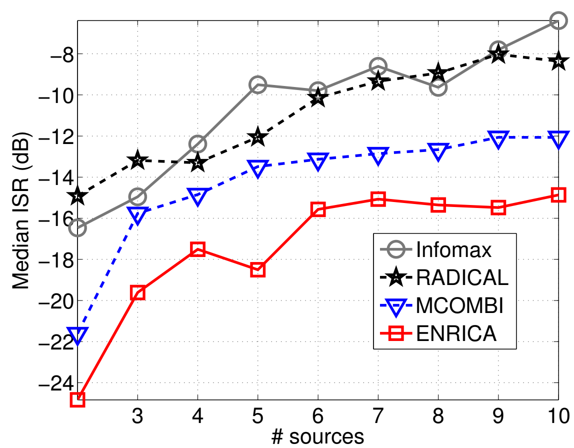

Fig. 2: 90th percentile of the Amari error (in dBs) across 100 random surrogates of mixtures of 2 to 10 EEG sources. Each source contained 5000 samples.

to its closer competitor (MCOMBI) is computation time. Separation of 10 sources with 5000 samples, required about 14 minutes for ENRICA compared to less than 1 second for MCOMBI [2]. However, this is acceptable in many EEG applications where reliability is far more important than computation time.

[2]The computations were all performed under MATLAB 7.8.0 (R2009a) for Windows XP, running on a Dell Optiplex 960 (Intel Core2 Quad CPU 2.83 GHz, 3.21 GB of RAM). All the code and datasets necessary for replicating the experiments shown in this article are available in the Internet [21].

V. Conclusions

A novel algorithm for blind separation of non-i.i.d. mutually independent sources has been introduced in this article. Its potential for the separation of brain sources underlying scalp EEG has been demonstrated using real EEG data. Moreover, our results cast some doubt on standard analysis practices among the neuroscientific community and suggest that BSS-based separation of many (e.g. more than 10) EEG sources might be unreliable in most experimental settings.

Acknowledgment

This work has been financially supported by the EU project GABA (FP6-2005-NEST-Path 043309).

References

[1] G. Gómez-Herrero, M. Atienza, K. Egiazarian, and J. L. Cantero, "Measuring directional coupling between EEG sources," *Neuroimage*, vol. 43, no. 3, pp. 497–508, 2008.

[2] Z. Koldovský, P. Tichavský, and E. Oja, "Efficient variant of algorithm fastica for independent component analysis attaining the Cramer-Rao lower bound," *IEEE T. Neural Networks*, vol. 17, no. 5, pp. 1265–1277, 2006.

[3] A. Yeredor, "Blind separation of Gaussian sources via second-order statistics with asymptotically optimal weighting," *IEEE Signal Proc. Let.*, vol. 7, no. 7, pp. 197–200, 2000.

[4] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of non stationary sources," *IEEE T. Signal Proces.*, vol. 49, no. 9, pp. 1837–1848, 2001.

[5] S. A. Cruces-Álvarez and A. Cichocki, "Combining blind source extraction with joint approximate diagonalization: Thin algorithms for ICA," in *Proc. ICA'03, Nara, Japan*, 2003, pp. 463–469.

[6] P. Tichavský, Z. Koldovský, A. Yeredor, G. Gómez-Herrero, and E. Doron, "A hybrid technique for blind separation of non-Gaussian and time-correlated sources using a multicomponent approach," *IEEE T. Neural Networks*, vol. 19, no. 3, pp. 421–430, 2008.

[7] P. Comon, "Independent component analysis - a new concept?" *Signal Proces.*, vol. 36, pp. 287–314, 1994.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[9] S. Hosseini and C. Jutten, "Markovian source separation," *IEEE. T. Signal Proces.*, vol. 51, no. 12, pp. 3009–3019, 2003.

[10] F. Vrins and M. Versleysen, "On the entropy minimization of a linear mixture of variables for source separation," *Signal Process.*

[11] S. A. Cruces-Alvarez, A. Cichocki, and S. Amari, "From blind signal extraction to blind instantaneous signal separation: criteria, algorithms, and stability," *IEEE T. Neural Networ.*, vol. 15, no. 4, pp. 859–873, 2004.

[12] L. F. Kozachenko and N. N. Leonenko, "A statistical estimate for the entropy of a random vector," *Probl. Infor. Transm.*, vol. 23, no. 95, p. 916, 1987.

[13] H. Stögbauer, A. Kraskov, S. A. Astakhov, and P. Grassberger, "Least dependent component analysis based on mutual information," *Phys. Rev. E 70*, vol. 6, 2004.

[14] T. Schneider and A. Neumaier, "Algorithm 808: ARFIT - a matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models," *ACM T. Math. Soft.*, vol. 27, no. 1, pp. 58–65, 2001.

[15] A. J. Bell and T. J. Sejnowski, “A non-linear information maximization algorithm that performs blind separation,” in *Advances in Neural Information Processing Systems 7*. Cambridge, MA: The MIT Press, 1995, pp. 467–474.

[16] E. G. Learned-Miller and J. W. Fisher III, “Independent Component Analysis using spacings estimates of entropy,” *J. Mach. Learn. Res.*, vol. 4, pp. 1271–1295, 2003.

[17] R. Boscolo, H. P. Pan, and V. P. Roychowdhury, “Independent component analysis based on nonparametric density estimation,” *IEEE T. Neural Networ.*, vol. 15, pp. 55–65, 2004.

[18] J.-F. Cardoso, “Blind signal separation: Statistical principles,” *P. IEEE*, vol. 86, no. 10, pp. 2009–2025.

[19] A. Vergult, W. De Clerq, A. Palmini, B. Vanrumste, P. Dupont, S. Van Huffel, and W. Van Paesschen, “Improving the interpretation of ictal scalp EEG: BSS-CCA algorithm for muscle artifact removal,” *Epilepsia*, vol. 48, no. 5, pp. 950–958, 2007.

[20] “EEGLAB user manual.” [Online]. Available: http://sccn.ucsd.edu/eeglab/maintut

[21] G. Gómez-Herrero and K. Rutanen, “MATLAB/C++ implementation of ENRICA,” 2009. [Online]. Available: http://www.cs.tut.fi/˜gomezher/projects/eeg/software.htm