This is the draft version of the article published in Communications in Computer Science (http://www.springer.com/series/7899), vol. 357, Springer (to appear).

APPLYING ICA IN EEG: CHOICE OF THE WINDOW LENGTH AND OF THE DECORRELATION METHOD

Gundars Korats[1,2], Steven Le Cam[1], Radu Ranta[1] and Mohamed Hamid[1]

[1] Université de Lorraine, CRAN UMR 7039, Vandoeuvre-les-Nancy, 54516, France
CNRS, CRAN UMR 7039, Vandoeuvre-les-Nancy, 54516, France
steven.le-cam@univ-lorraine.fr, radu.ranta@univ-lorraine.fr
[2] Ventspils University College, 101 Inzenieru iela, LV-3601, Ventspils, Latvia
gundars.korats@gmail.com

Abstract. Blind Source Separation (BSS) approaches for multi-channel EEG processing are popular, and in particular Independent Component Analysis (ICA) algorithms have proven their ability for artefacts removal and source extraction for this very specific class of signals. However, the blind aspect of these techniques implies well-known drawbacks. As these methods are based on estimated statistics from the data and rely on an hypothesis of signal stationarity, the length of the window is crucial and has to be chosen carefully: large enough to get reliable estimation and short enough to respect the rather non-stationary nature of the EEG signals. In addition, another issue concerns the plausibility of the resulting separated sources. Indeed, some authors suggested that ICA algorithms give more physiologically plausible results than others. In this paper, we address both issues by comparing four popular ICA algorithms (namely FastICA, Extended InfoMax, JADER and AMICA). First of all, we propose a new criterion aiming to evaluate the quality of the decorrelation step of the ICA algorithms. This criterion leads to a heuristic rule of minimal sample size that guarantees statistically robust results. Next, we show that for this minimal sample size ensuring constant decorrelation quality we obtain quasi-constant ICA performances for some but not all tested algorithms. Extensive tests have been performed on simulated data (i.i.d. sub and super Gaussian sources mixed by random mixing matrices) and plausible data (macroscopic neural population models placed inside a three layers spherical head model). The results globally confirm the proposed rule for minimal data length and show that the use of sphering as decorrelation step might significantly change the global performances for some algorithms.

Keywords: EEG : BSS : ICA : whitening : sphering

1 INTRODUCTION

The analysis of electro-physiological signals generated by brain sources leads to a better understanding of brain structures interaction and is useful in many clinical applications

or for brain-computer interfaces (BCI) [1]. One of the most commonly used method to collect these signals is the scalp electroencephalogram (EEG). The EEG consists in several signals recorded simultaneously using electrodes placed on the scalp (see fig.1). The electrical activity of the brain sources is propagated through the anatomical structures and the resulting EEG is a linear mixture (with unknown or difficult to model parameters) of brain sources and other electro-physiological disturbances, often with a low signal to noise ratio (SNR) [2]. It is widely assumed that electrical brain potentials recorded by the electrodes mainly arise from synchronous activity of neurons within localized cortical *patches*. The far-field projection of such locally generated activity can be suitably model by the projection of a single equivalent current dipole placed at the center of the patch, resulting in a linear mixing of mostly dipolar sources on the EEG [3].

The blind source separation (BSS) is a nowadays well established method to retrieve original sources from the EEG mixing, as it can estimate both the mixing model and original sources [4]. In particular, approaches based on High Order Statistics (HOS) such as Independent Component Analysis (ICA) are common methods in this context and have been very useful for denoising purpose or brain sources identification. Generally, ICA algorithms include a preliminary decorrelation step based on second order statistics (estimated on a user chosen window length), which serves as an initialization for the next optimization step (independence maximization). Still, there is an infinite number of possible decorrelation matrices (as they are determined up to an arbitrary rotation). The two most popular decorrelation techniques are whitening and sphering, and it seems that they might influence the final separation results, especially in EEG applications [5]. In this paper, two issues will then be evaluated: 1) the accuracy of the decorrelation matrix estimation given the considered data length and 2) the sensitivity to the initialization step using whitening or sphering in the specific context of dipolar sources mixing. Four ICA algorithms based on HOS have been chosen: FastICA [6], Extended InfoMax [7], AMICA [8] and JADER [9].

1) The use of BSS on EEG signals implicitly assumes that the estimated second order statistics are meaningful. In order to ensure the reliability of these statistics, different authors propose optimal sample sizes (i.e. EEG signal time points), generally equal to $k \times n^2$ where n is number of channels and k is some empirical constant varying from 5 to 32 [10,11,12]. If these assumptions are correct, large amount of channels requires huge sample sizes, processing and time resources. On the other hand, EEG signals are at most short term stationary, so it would be interesting to find a sufficient inferior bound for the number of necessary samples. The first question is then how to define a minimum sample size that provides reliable estimation of sources and mixing model.

2) The second issue addressed in this paper concerns the sensitivity of the BSS/ICA performance given the initial decorrelation step in the dipolar mixing context. In the literature [13], some authors observed that using different initializations (different decorrelation methods like classical whitening or sphering), the results are more or less biologically plausible, meaning that more or less dipolar sources are retrieved from the data. A recent extensive study from the same authors [5] proposed an evaluation of the ability of 18 source separation methods to result in maximally independent com-

ponent processes with nearly dipolar scalp projection. The results show that AMICA and Extended InfoMax give better performances compared to FastICA and JADER. Both AMICA and Extended InfoMax begin by sphering the data, while FastICA and JADER begin with a classical whitening step. We would like to evaluate the impact of whitening and sphering on these four ICA algorithm performances. Unlike the previous studies that are directly using real EEG data, this evaluation will be performed on simulated data, giving the possibility of a controlled quantification of the algorithm performances. The evaluation is here proposed in the context of randomly generated data using i.i.d. sub and super Gaussian sources mixed by random mixing matrices, and in the context of plausible EEG data generated by macroscopic neural population models placed inside a three layers spherical head model.

The paper is organized as follow: section 2 exposes the EEG forward problem, explains the basics of the BSS methodology and gives some details on the four evaluated ICA methods. Section 3 proposes a normalized Riemannian likelihood as an evaluation criterion for the accuracy of the covariance matrix estimation and recalls the separability performance index used to evaluate the ICA algorithms. In section 4 both random and biologically plausible data set are described, while estimation and separation performances of the algorithms facing these data set are provided in section 5. In section 6, these results are discussed and future works are considered.

2 PROBLEM STATEMENT

2.1 EEG mixing model

Classical EEG generation and acquisition model is presented in Figure 1. It is widely accepted that the signals collected by the sensors are linear mixtures of the sources [2].

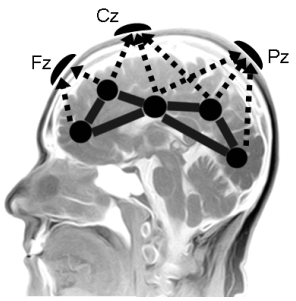


Fig. 1. EEG linear model

Subsequently, the EEG mixture can be written as

$$\mathbf{X} = \mathbf{AS}, \tag{1}$$

where $\mathbf{X}$ are the observations (electrodes), $\mathbf{A}$ is the mixing system (anatomical structure) and $\mathbf{S}$ are the original sources.

2.2 EEG separation model

We restrain in this paper to classical well determined mixtures, where the number of channels is equal to the number of underlying sources. In this case, BSS gives the linear transformation (separating) matrix $\mathbf{H}$ and the output signal vector $\mathbf{Y} = \mathbf{HX}$, containing source estimates. Ideally, the global system matrix $\mathbf{G} = \mathbf{HA}$ between the original sources $\mathbf{S}$ and their estimates $\mathbf{Y}$ will be a permuted scaled identity matrix, as it can be proven that the order and the original amplitude of the sources cannot be recovered [4].

In almost all BSS methods, the matrix $\mathbf{H}$ is obtained as a product of two statistically based linear transforms: $\mathbf{H} = \mathbf{JW}$ with

- $\mathbf{W}$ performing data orthogonalization: whitening/sphering,
- $\mathbf{J}$ performing data rotation : independence maximization via higher-order statistics (HOS) or joint decorrelation of several time (frequency) intervals

The first step (data decorrelation) can be seen as an initialization for the second step. In theory any orthogonalization technique can be used to initialize the second step but in this paper we will focus on two popular decorrelation techniques: whitening (classical solution) and sphering (assumed to be more biologically plausible [13]).

BSS initialization: whitening/sphering

Whitening In general EEG signals $\mathbf{X}$ are correlated so their covariance Σ will not be a diagonal matrix and their variances will not be normalized. Data whitening means projection in the eigenspace and normalisation of variances. The whitening transform can be computed from the eigen-decomposition of the data covariance matrix $\Sigma = \Phi\Lambda\Phi^T$:

$$\mathbf{X_w} = \Lambda^{-\frac{1}{2}}\Phi^T\mathbf{X}, \tag{2}$$

where Λ and Φ are the eigenvalues and the eigenvectors matrices respectively. After (2), the signals are orthogonal and with unit variances (Figure 2(c)).

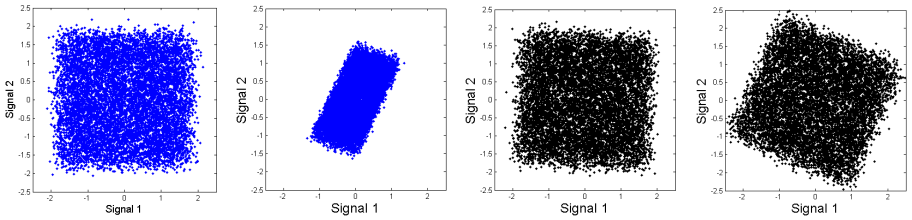
Sphering completes whitening by rotating data back to the coordinate system defined by principal components of the original data [14]. In other words, sphered data are turned as close as possible to the observed data (Figure 2(d)):

$$\mathbf{X_{sph}} = \Phi\Lambda^{-\frac{1}{2}}\Phi^T\mathbf{X}. \tag{3}$$

2.3 Optimization: data rotation

Second step would be finding a rotation matrix $\mathbf{J}$ to be applied to the decorrelated data (whitened or sphered) in order to maximize their independence. Rotation can be done using second order statistics (SOS) using joint decorrelations and/or using HOS cost functions. We restrain here to the second (HOS) approach[3]. Several cost functions

[3] As described in the next section, in our simulations we used random non-Gaussian stationary data, without any time-frequency structure. Therefore algorithms based on SOS as SOBI, SOBI-RO and AMUSE were not used.



(a) Original sources (b) Mixed data (c) Whitened data (d) Sphered data

Fig. 2. Example of different decorrelation approaches for two signals

and optimization techniques were described in the literature (see for example [4,12]). Among the most well known and used in EEG applications, we can cite FastICA (negentropy maximization [6]), Extended InfoMax (mutual information minimization [7]) and JADER (joint diagonalization of fourth order cumulant matrices [9]). Another recent algorithm has been proposed by Palmer *et al.* and is called AMICA [8]. Based on the modeling of each source component as a sum of extended Gaussians, this method has shown very promising results in the context of EEG data [5].

Specifically, in this paper we test the performances and the robustness of these four ICA algorithms with respect to the sample size and the initialization step in both contexts of random unstructured data mixing and biologically plausible data mixing.

3 Performance evaluation criteria

3.1 Reliable estimate of the covariance: Riemannian likelihood

As noted before, BSS model consists of decorrelation and rotation. Both steps are based on statistical estimates. The first step is common for all algorithms and relies on the estimation of the covariance matrix. Therefore it is necessary to have reliable estimates of this matrix. In other words, given a known covariance matrix Σ, we want to evaluate the minimum sample size N necessary to obtain a covariance matrix estimation $\hat{\Sigma}_N$ close enough to the original one with respect to a distance that we have to define.

We propose here an original distance measure between the true and the estimated covariance matrices, inspired from digital image processing and computer vision techniques [15]. In the context of object tracking and texture description, a distance measure is used to estimate whether an observed object or region corresponds to a given covariance descriptor. To estimate similarity between matrices respectively corresponding to the target model and the candidate, and knowing that covariance matrices are symmetric positive definite, the following general[4] distance measure can be used:

$$d^2(\hat{\Sigma}_N, \Sigma) = \mathrm{tr}\left(\log^2\left({\hat{\Sigma}_N}^{-\frac{1}{2}}\Sigma{\hat{\Sigma}_N}^{-\frac{1}{2}}\right)\right) \tag{4}$$

[4] on Riemannian manifolds

In the ideal case of a perfect estimation, the matrix $\mathbf{C} = \hat{\Sigma}_N^{-\frac{1}{2}} \Sigma \hat{\Sigma}_N^{-\frac{1}{2}}$ equals the identity matrix $\mathbf{I}_n$ and d becomes 0 (n being the number of measured signals, equal to the source number in our case). In real cases though, assuming that the covariance estimate is not very far from the real covariance matrix, $\mathbf{C} = \mathbf{I}_n + \varepsilon$, ε being a symmetric error matrix. In this case, using the eigenvalues decomposition and the properties of the trace of a symmetric matrix, equation (4) can be rewritten as:

$$d^2(\hat{\Sigma}_N, \Sigma) = \mathrm{tr}\left(\log^2(\mathbf{I}_n + \varepsilon)\right) \tag{5}$$

$$= \mathrm{tr}\left(\mathbf{U}\log^2(\mathbf{D}_{\mathbf{I}_n+\varepsilon})\,\mathbf{U}^T\right) \tag{6}$$

$$= \mathrm{tr}\left(\log^2(\mathbf{D}_{\mathbf{I}_n+\varepsilon})\right) \tag{7}$$

$$\approx \Sigma_{i=1}^{n} \log^2(1+\varepsilon_{ii}) \tag{8}$$

Now, using the fact that for small ε, $\log(1+\varepsilon) \approx \varepsilon$ and assuming that the errors are equally distributed over the diagonal $\mathbf{D}_{\mathbf{I}_n+\varepsilon}$, (5) becomes:

$$d^2(\hat{\Sigma}_N, \Sigma) \approx n\varepsilon^2 \tag{9}$$

proportional with the matrix dimension n and, in our case, with the number of EEG channels. In order to avoid this channel number effect, we propose to modify the distance by multiplying it by k/n, with k being a user chosen constant ensuring the desired level of the estimation quality:

$$d_n{}^2(\hat{\Sigma}_N, \Sigma) = \frac{k}{n}\mathrm{tr}\left(\log^2\left(\hat{\Sigma}_N^{-\frac{1}{2}} \Sigma \hat{\Sigma}_N^{-\frac{1}{2}}\right)\right) \tag{10}$$

As in [15], we adopt an exponential function of the modified distance d_n as the local likelihood

$$p(\Sigma_N) \propto exp\{-\lambda \cdot d_n{}^2(\Sigma, \hat{\Sigma}_N)\}. \tag{11}$$

with the parameter λ fixed to the constant value $\lambda = 0.5$ [15]. This $p(\Sigma_N)$ value varies between 0 and 1, 1 meaning perfect estimation ($\Sigma = \hat{\Sigma}_N$). A $p(\Sigma_N)$ value of 0.95 is considered as a well chosen threshold above which the covariance matrices are considered to be approximately equal.

3.2 Separability Performance Index

In order to measure the global performance of BSS algorithms (orthogonalization plus rotation), we use the performance index (PI) [4] defined by

$$PI = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \frac{|g_{ij}|}{\max_k |g_{ik}|} - 1 \right) + \frac{1}{2n(n-1)} \sum_{j=1}^{n} \left(\sum_{i=1}^{n} \frac{|g_{ij}|}{\max_k |g_{kj}|} - 1 \right) \tag{12}$$

where $\mathbf{g}_{ij}$ is the (i,j)-element of the global $n \times n$ system matrix $\mathbf{G} = \mathbf{HA}$, $\max_k |g_{ik}|$ is the maximum value among the absolute values of the elements in the ith row of $\mathbf{G}$ and $\max_k |g_{kj}|$ is the maximum value among the absolute values of the elements in the jth column of $\mathbf{G}$. Perfect separation yields a null performance index. In practice a PI under 10^{-1} means that the separation result is reliable.

4 Simulated data set

The algorithms performances are assessed on two types of data. Following our BIOSIGNALS paper [16], the first data set consists in simulated generalized Gaussian sources mixed by randomly simulated matrices. The second one is obtained by mixing sources given by macroscopic neural populations models [17] with mixing matrices computed from a realistic three layers lead field model.

4.1 Random data set

We have chosen to simulate stationary white source signals, as the retrieval of time structures is not the purpose of this work (in fact, in all the tested algorithms, as in most of the HOS type methods, the time structure is ignored). In order to simulate sources with realistic probability distributions, we analysed depth intra-cerebral measures (SEEG). According to our observations (see also [11,10]), the probability distribution of the electrical brain activity signals can be suitably modelled by Generalized zero-mean Gaussians, as shown in fig. 3(a) and fig. 3(b)). For this reason we used randomly generated both supergaussian (Laplace - Figure 3(c)) and subgaussian (close to uniform (Figure 3(d))) distributions.

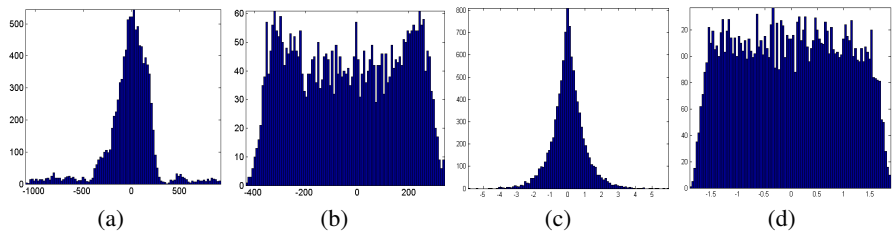


Fig. 3. Histograms of real SEEG samples ((a) background and (b) ictal activities) and histograms of generalized Gaussian simulated data ((c) super-Gaussian and (d) sub-Gaussian data).

Several simulations were made, using 8, 16, 24, 32 and 48 source signals. Half of the sources were generated as supergaussian and half as subgaussian. The sources were afterwards mixed using a randomly generated mixing matrix $\mathbf{A}$ (uniform distribution in $[-1,1]$). We then consider here the performance of each of the four ICA algorithms facing simulated stationary non-artefacted data. Such evaluation is likely to give us a rule of a minimum amount of data needed for a reliable source separation in favourable conditions (after an artefact elimination step for example).

4.2 Plausible data set

More realistic contexts is to be simulated in order to evaluate the behaviour of the algorithm confronted to the real EEG BSS problem. We then propose a second data

set where the sources are simulated by a macroscopic model [17] able to reproduce normal background activity as well as pre-ictal and ictal (epileptic) electromagnetic activity. The mixing matrices are designed using a fast accurate three layers head model [18]. Such mixing matrices guarantee the dipolarity of the underlying sources to be retrieved, providing a framework where the influence issue of the initialization (whitening or sphering) in such context can be addressed.

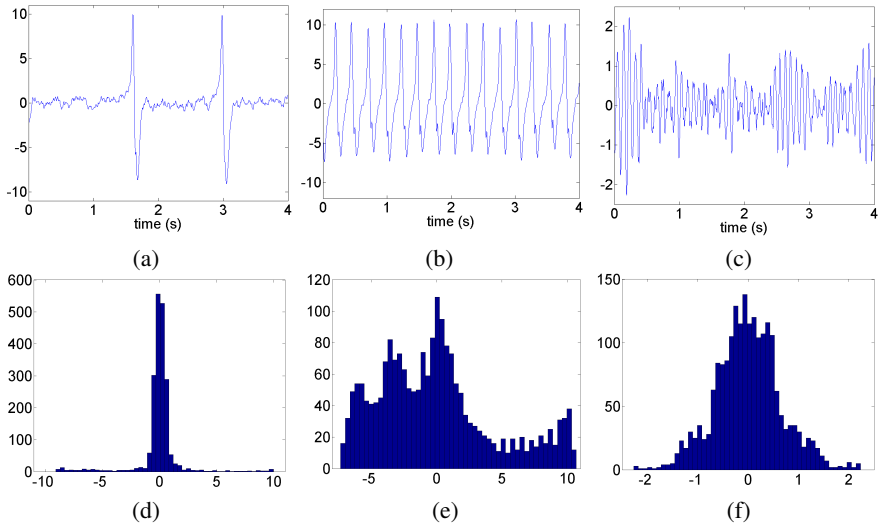


Fig. 4. Realistic activities and corresponding histograms. first pre-ictal: (a) & (d), second pre-ictal: (b) & (e), second ictal: (c) & (f).

Realistic sources Macroscopic models describe the neuronal activity at the scale of neuronal populations by modelling the interconnection of pyramidal cells with inhibitory or excitatory inter-neurons. They have been particularly used to successfully generate realistic electrophysiological recordings [19,20]. In this work we have chosen the Wendling's model [17], described by a set of ten differential equations. It has been shown that this model is able to reproduce normal background activity (inter-ictal), first and second pre-ictal activities as well as first and second ictal activities. Parameter values to be chosen in order to get these distinct epileptic activities are detailed in [21]. In this paper, these simulated activities have been introduced as sources (see fig. 4.2), excepted the normal background and the first ictal activities that have Gaussian-like distribution and are then inadequate for the selected ICA algorithms.

In a realistic situation, the number of ictal sources to be retrieved is limited. In our simulation, the number of realistic sources (from pre-ictal to ictal) has been chosen to be an eighth of the number of channels n (from 1 for 8 channels to 6 for 48 channels).

The remaining background activities are simulated randomly as sub and super gaussian like in the previous random data set.

Dipolar mixing matrix The next step for obtaining realistic EEGs, after plausible source generation, is the construction of a realistic mixing matrix. We obtained it using the classical three spheres model of Rush & Driscoll and the Berg surface potential estimation method [18]: the realistic sources generated as described in the previous subsection were assumed to be time courses of brain dipoles. The positions and the orientations of these dipoles (8 to 48) were randomly generated inside the inner sphere of a the head model, representing the brain. The six parameters of the dipoles being thus completely defined (Cartesian coordinates, angles and magnitude), one can generate the corresponding scalp map (i.e. the electrical potentials on the outer sphere) using the so-called Lead Field matrix corresponding to the three-spheres forward model (see for example [22]). We used here the Berg technique [18], which accurately and rapidly approximates the scalp potentials generated by a dipole for a three spheres model with the weighted sum of the potentials generated by three dipoles in a single sphere model.

The Lead Field matrix allows the computation of the scalp map (potentials) for every point on the surface of the outer sphere. Still, in the EEG simulation context, we are only interested by the potentials recorder by the scalp electrodes. In our simulation, we used as a basis montage the classical 10-10 EEG montage (figure 5). We have next chosen five subsets consisting of 8, 16, 24, 32 and 48 electrodes corresponding to real EEG applications (sleep studies, brain computer interfaces and clinical setups for epilepsy diagnostics).For example, for the 8-electrodes montage, the chosen electrodes were $F_{pz}, T_7, C_3, C_z, C_4, T_8, O_1$ and O_2; for the 16-electrodes montage, $F_3, F_z, F_4, T_7, C_3, C_z, C_4, T_8, P_7, P_3, P_z, P_4, P_8, PO_7, PO_8, O_z$; and son on.

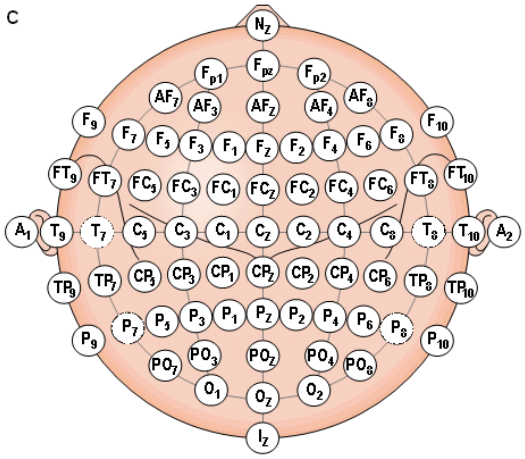


Fig. 5. Electrode placement for the 10-10 montage.

5 RESULTS AND DISCUSSION

The four algorithms are first evaluated on a random data set in order to define a minimum data length rule ensuring accurate separation performances and to analyse the impact of initialization step in the general toy case of unstructured randomly mixed data. An equivalent study is then applied on a more physiologically plausible data set, on which our minimum data length rule is validated. On the same plausible data, the results are evaluated in order to confirm or infirm the superiority of sphering initialization over whitening initialization in the context of dipolar sources separation. For each data set, the four algorithms are evaluated on 8, 16, 24, 32 and 48 source sizes with sample size varying from 1s to 20s with a 1s step (at a 512 Hz sampling rate). The number of iteration for each source size/sample size has been set to 50, a new set of data (random data sources, plausible data sources, mixing matrices) being simulated at each iteration.

5.1 Random data set

A minimum length rule This section presents the results of the covariance estimation accuracy vs the length of the data. In a previous work [16], the distance between covariance matrices was computed using (4) from different sample sizes starting from 100 till 5000 by 100 points step, and number of channels taking values in the set $\{8, 10, 12, 14, 16, 18\}$. The likelihood was further evaluated using (11). A constant threshold was empirically fixed to $p = 0.95$ (Figure 6(a)): likelihood values above this threshold was assumed to guarantee good estimation of covariances as stated in section 3.1. However, this previous study already outlined that this rule might be too strong, leading in a ever decreasing *PI* with the number of channels. This observation has been confirmed when studying this evolution of *PI* on larger number of channels. Therefore,we proposed the *normalized* Riemannian distance defined by (10) with the objective to refine our minimum length rule. Again, likelihood values above a threshold of 0.95 were assumed to guarantee good estimation of second order statistics (Figure 6(a)). The normalization factor k has been experimentally set to 8, taking the 8 channels mean error ε as a reference on which higher number of channels configurations are scaled.

This rule is reported on the figure 6(c). The proposed rule is rather linear, thus being in contradiction with current literature suggestions, rather proportional to n^2. Our rule is then between the bounds given in the literature for low number of channels, but is increasing much slower and gives lower bounds for number of channels above 24. A possible way to interpret the figure 6(c) is to use it as a decision rule: for a given number of channels, one can estimate the minimum number of data points necessary to have a reliable estimate of the covariance matrix and thus a reliable whitening. This decision rule leads to data lengths between approximately 2s (1024 data points) for 8 channels to 7s (3584 data points) for 48 channels. This range of time length is more compatible with the stationarity hypothesis than the values obtained using the $30n^2$ rule [12,11]. Indeed, with this rule, we get from 1920 (3.75s) to 9720 (2min15s) data points respectively for 8 and 48 channels, which is rather contradictory (at least in a realistic EEG setup) with the assumption of stationarity on which most of BSS/ICA algorithms are based[5].

[5] This observation is important especially for high resolution EEGs, having a high number of channels.

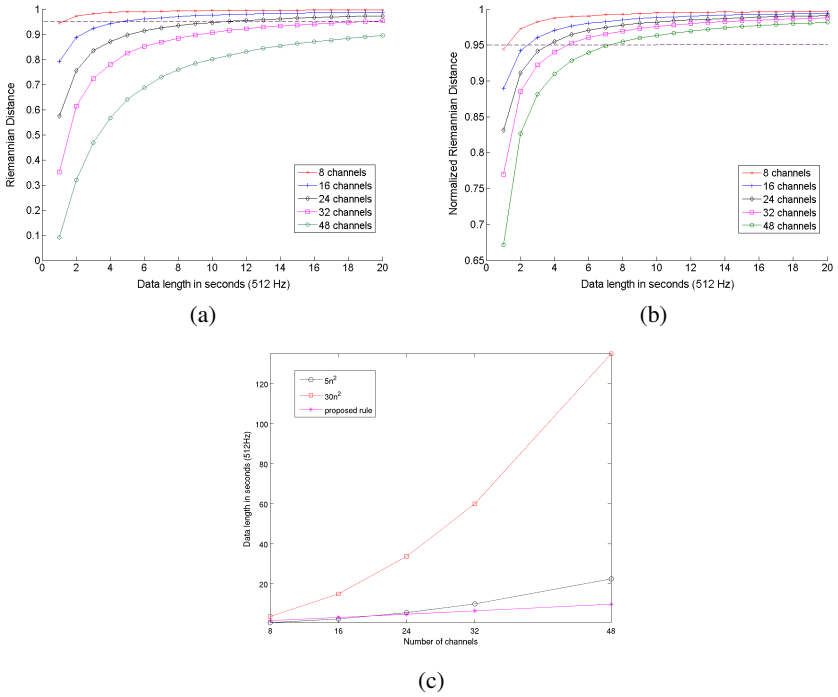


Fig. 6. Results on the random data set: (a) Riemannian Distance, (b) *normalized* Riemannian Distance and (c) Minimum length rules from literature vs proposed (linear) rule.

In order to experimentally validate this length rule, we computed the performance index *PI* (12) for the resulting data length. Mean (over 50 realizations) of the *PI* as well as its standard deviation for each algorithm (with whitening and sphering initializations) are reported in the Table 1. As it can be seen, FastICA (with either whitening or sphering) gives rather stable *PI* under 0.05 for this given rule (from 0.048 for 8 channels to 0.038 for 48 channels). It has to be noticed that *PI* values indicate better performances when the number of channels is increasing with respect to our empirical rule. This could suggest that our proposed criterion could be relaxed and the number of points could be reduced further for FastICA. On the other hand, one must take into account that these tests are performed on simulated stationary random data: if outliers are present, HOS estimates are more affected than the SOS estimations used to define our threshold, thus a higher amount of points might be needed for HOS reliable estimation.

This observation holds when it comes to AMICA and JADER for number of channels from 8 to 32, the rule being not verified for 48 channels in this particular case of random data set. A quick look at the figure 5.1(e) (*PI* vs sample size for 48 channels) shows a rupture of the JADER curves around 6s to 8s, 9s being required to get a *PI* under 0.1. Such rule seems thus to be inadequate for JADER algorithm for number of channels over 48.

Table 1. Perfomance index (PI) values for random mixtures: mean and standard deviation for the four algorithms, with two initializations (whitening and sphering) and using the couple n/N (nb. of channels / data length) given by the heuristic rule derived from figure 5.1(a).

	$n=8$ $N=2\times512$		$n=16$ $N=3\times512$		$n=24$ $N=4\times512$		$n=32$ $N=5\times512$		$n=48$ $N=7\times512$	
	W	S	W	S	W	S	W	S	W	S
FastICA	0.048	0.049	0.043	0.043	0.041	0.040	0.040	0.040	0.038	0.038
	(0.014)	(0.012)	(0.006)	(0.007)	(0.006)	(0.006)	(0.005)	(0.005)	(0.004)	(0.006)
AMICA	0.024	0.029	0.018	0.019	0.021	0.029	0.036	0.061	0.113	0.170
	(0.011)	(0.031)	(0.004)	(0.005)	(0.021)	(0.051)	(0.048)	(0.056)	(0.050)	(0.060)
Extended	0.167	0.199	0.274	0.326	0.288	0.334	0.252	0.300	0.274	0.310
Infomax	(0.091)	(0.124)	(0.059)	(0.055)	(0.022)	(0.013)	(0.019)	(0.014)	(0.008)	(0.007)
JADER	0.044	0.044	0.038	0.038	0.035	0.035	0.035	0.035	0.130	0.132
	(0.010)	(0.010)	(0.005)	(0.005)	(0.004)	(0.004)	(0.003)	(0.003)	(0.029)	(0.032)

Extended Infomax is showing bad performances for these data length, requiring much more sample size to converge, confirming the results presented in [23]. This phenomenon appears because of our choice of the simulated data. Indeed, because of the use of subgaussian sources, the algorithm (even in the extended version) needs more data points in order to give reliable results. Empirically, one can say that the $30n^2$ rule seems to be adapted for the Extended InfoMax algorithm, but apparently too strong for the three others.

In the case of AMICA, the initialization parameter has to be considered in order to fully understand its misadequation with our minimum data length rule in the random data set case for 48 channels (see below the analysis for the plausible data set).

Impact of initialization Our second objective is to analyse the sensitivity of the ICA algorithms to the initialization step with whitening or sphering. The curves of figure 5.1 are showing evolution of *PI* with the data sample size for the five channel number configurations considered. Whitening and sphering curves are difficult to distinguish for FastICA and JADER, allowing to conclude that these methods are not sensitive to the decorrelation step. This has to be explained by the optimization strategy of these methods, based respectively on a fixed point and a Jacobi technique, both techniques reputed to have fastest convergence and being more reliable than the gradient technique [9]. AMICA is doing better than FastICA and JADER in most configurations when the amount of data is enough. This algorithm is based on the fitting of extended Gaussian (mixtures of scaled Gaussians) for each source time course, thus needing more data and execution time for accurate estimation and convergence (see the note below on time convergence). Besides, results appear to be quite deceiving for AMICA when it comes to the 48 channels configuration, with a *PI* around 0.06 for lengths greater than 10s for whitening (with a large standard deviation around 0.05), but a *PI* well above 0.1 for sphering. In this case initialization shows to have a noticeable impact on AMICA. This observation can be done also for Extended InfoMax for all five channel size cases. For this specific data set of randomly mixed sources, whitening initialization (solid curves) is resulting globally in better *PI* than sphering initialization (dashed curves).

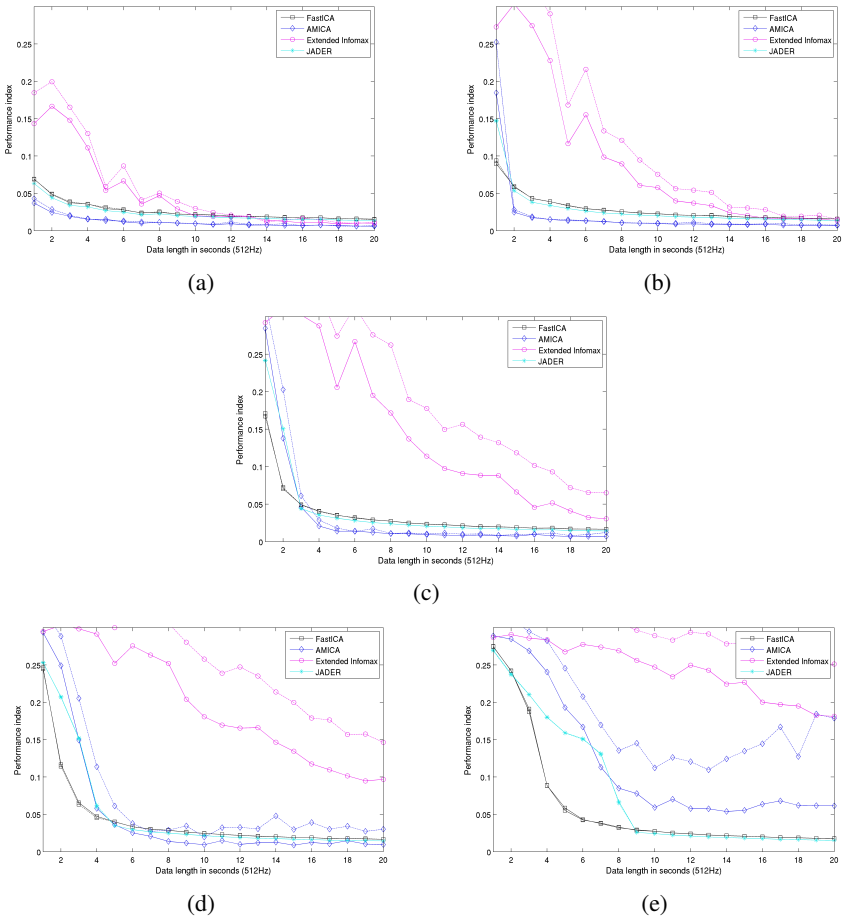


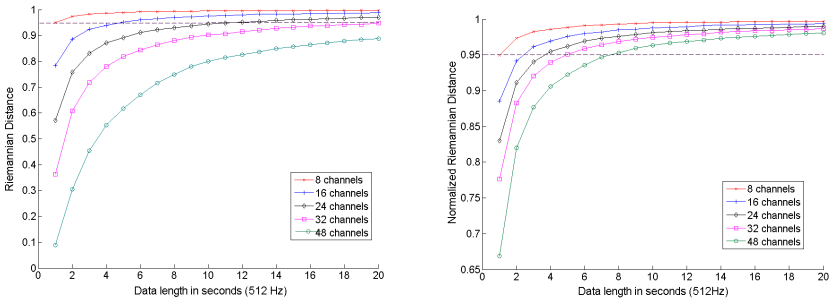
Fig. 7. Results on the random data set: performance index (*PI*) curves vs data length for (a) 8 channels, (b) 16 channels, (c) 24 channels, (d) 32 channels and (e) 48 channels. Initialization with whitening (solid lines) and sphering (dashed lines).

As pointed out in [5], Extended InfoMax and AMICA are based on a natural gradient descent optimization scheme, initialization is then a major issue for these algorithms: the farthest from the solution the initialization is, the longest will take the optimization procedure. In the case of random mixing matrices, the solutions are distributed widely over the optimization space, making it difficult to define an adequate initialization point. In this context, whitening seems to be on average more appropriate than sphering. It has to be noticed that in our simulation no iteration or convergence criteria parameter has been changed in the Extended InfoMax algorithm, while maximum number of iterations has been set to 300 for the AMICA procedure (some numerical issues have been experienced with the default 100 value for short length data (<=2s)).

Note on time convergence: Due to the large amount of parameters to be estimated by AMICA, it might be important to notice that this method is extremely time consuming compared to JADER and FastICA, and also much slower than Extended InfoMax. To give an idea: while FastICA is taking less than 1s on 32 channels and 20*s* data length (mean time observed for 50 iterations on random data), JADER requires no more than 3s, Extended InfoMax needs up to 3 minutes, and AMICA requires almost 4 minutes. On the other hand, AMICA is *per se* much more flexible than Extended InfoMax which only try to fit a single generalized Gaussian distribution on each source, explaining the better performances of AMICA when compared to Extended InfoMax.

5.2 Plausible data set

Minimum Length Rule validation As it can be observed by comparing the computed (normalized) Riemannian distance, covariance estimations on this data set (figure 5.2) show to be very similar to the results obtained on the previous random data set. Thus, plausible source time courses and mixture do not show to have high influence on these second order statistics estimation, allowing to keep the same minimum data length rule derived from figure 6(c). Table 2 gives mean *PI* related to this proposed decision rule on the plausible data set, where it can be seen that these minimum data length bounds appear to be adequate for all the algorithms in most channel size configurations, even for Extended InfoMax when a sphering initialization is considered. Some lower performances are observed for JADER for 48 channels, confirming the observation made in the previous section that this decision rule might be inadequate for this method for high number of channels. Relative better performances observed on Extended InfoMax, and over all impressive results given by AMICA with sphering for channel size over 24 (mean $PI < 0.015$ with minimal standard deviation of 0.001) have to be explained in the light of the initialization parameter.



(a) (b)

Fig. 8. Results on the plausible data set: (a) Riemannian Distance and (b) *normalized* Riemannian Distance.

Table 2. Perfomance index (PI) values for plausible EEG: mean and standard deviation for the four algorithms, with two initializations (whitening (W) and sphering (S)) and using the couple n/N (nb. of channels / data length) given by the heuristic rule derived from figure 5.1(a).

	$n=8$ $N=2\times512$		$n=16$ $N=3\times512$		$n=24$ $N=4\times512$		$n=32$ $N=5\times512$		$n=48$ $N=7\times512$	
	W	S	W	S	W	S	W	S	W	S
FastICA	0.048 (0.014)	0.047 (0.012)	0.044 (0.007)	0.044 (0.008)	0.038 (0.006)	0.038 (0.006)	0.036 (0.004)	0.037 (0.004)	0.034 (0.004)	0.033 (0.004)
AMICA	0.024 (0.012)	0.023 (0.009)	0.019 (0.004)	0.018 (0.003)	0.015 (0.001)	0.015 (0.002)	0.015 (0.008)	0.013 (0.001)	0.020 (0.021)	0.011 (0.001)
Extended Infomax	0.159 (0.099)	0.089 (0.046)	0.207 (0.061)	0.097 (0.038)	0.218 (0.038)	0.085 (0.026)	0.162 (0.024)	0.055 (0.015)	0.173 (0.017)	0.058 (0.011)
JADER	0.040 (0.008)	0.040 (0.008)	0.039 (0.006)	0.039 (0.006)	0.036 (0.004)	0.036 (0.004)	0.037 (0.003)	0.037 (0.003)	0.123 (0.032)	0.123 (0.032)

Sphering is better than whitening for dipolar sources separation Figure 5.2 displays the evolution of *PI* with the data sample size for the five considered configurations (channel number). A quick look at these curves let us conclude that FastICA and JADER are unsensitive to their initialization as expected and explained in the previous section. No major differences on the performances can be noticed between the random data set and the plausible data set, confirming the reputation of stability and reliability of these techniques in various situations. Concerning natural gradient descent based algorithms (Extended InfoMax and AMICA), the behaviour changes radically from the first data set to the second one. Results improve for both methods, especially when a sphered initialization is used. Extended InfoMax show convergence with *PI* under 0.1 in the five configurations when data is used. AMICA is found out to show very good performance facing mixtures of dipolar sources, with a high robustness to low sample sizes for high number of channels, especially when initialized with sphering: 3s appears to be enough to get a *PI* under 0.03 for 48 channels, with a standard deviation of 0.01.

In this particular case of dipolar mixing, the reasons of the superiority of sphering over whitening can be found in [5]: *The objective of Principal Component Analysis (PCA)*[6] *is to lump together as much variance as possible into each successive principal component, whose scalp maps must then be orthogonal to all the others and therefore are not free to model a scalp source projection resembling a single dipole.* In other words, whitening initialize the algorithm far from the solution, leading to a more difficult convergence for methods based on natural gradient descent like Extended InfoMax and AMICA. On the other hand, Delorme et al. [5] emphasize that: *Sphering components, in particular, most often have stereotyped scalp maps consisting of a focal projection peaking at each respective data channel and thus resembling the projection of a radial equivalent dipole.*. Consequently, sphering leads to initialization point much more closer to the solution than whitening in the dipolar case, as it is confirmed and quantified by our results.

[6] equivalent to whitening

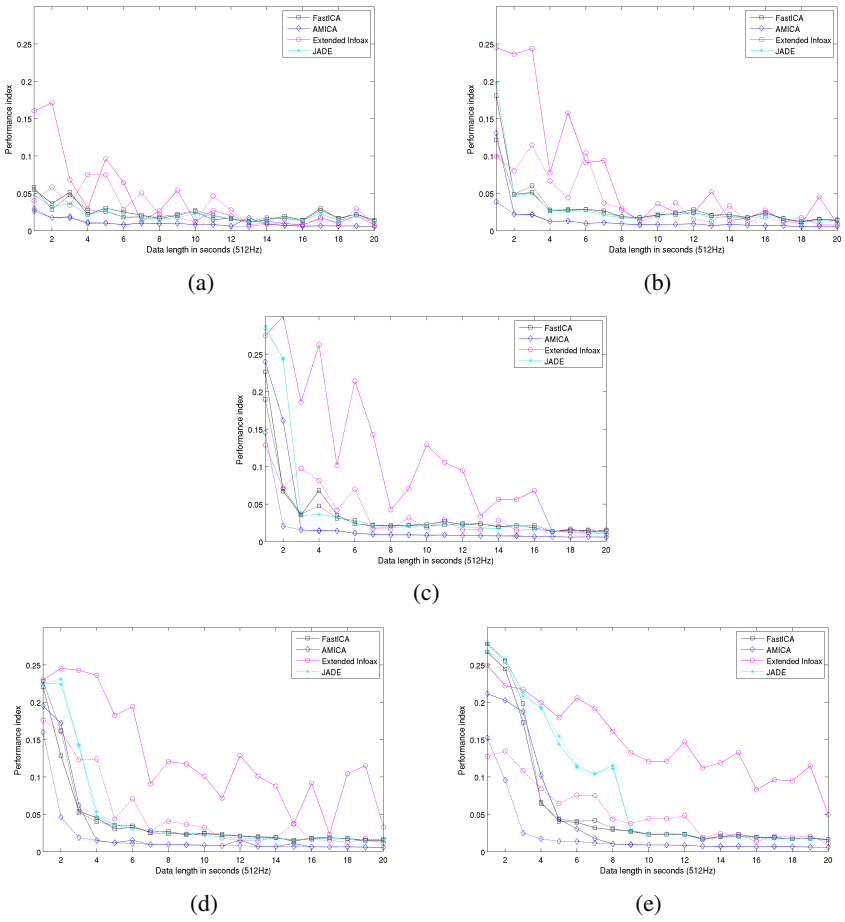


Fig. 9. Results on the plausible data set: performance index (*PI*) curves vs data length for (a) 8 channels, (b) 16 channels, (c) 24 channels, (d) 32 channels and (e) 48 channels. Initialization with whitening (solid lines) and sphering (dashed lines).

6 Conclusions and Future work

The first goal of this paper was to define a low bound of data length for robust separation results. Four ICA algorithms often used to analyse EEG signals were tested on different data lengths (1 to 20 seconds at 512 Hz sampling rate) and number of signals (8, 16, 24, 32 and 48 sources/channels). A rule of minimum sample size is derived from separation results on a random data set consisting of subgaussian and supergaussian source signals mixed by random mixing matrices, and is validated on a plausible data set in which sources were simulated by a macroscopic model of neuronal population and mixed by dipolar mixing matrices obtained from a three layers head model.

This low bound is based on an original, normalized distance measure inspired by the computer vision community and leads to a reasonable minimum time length. According to our results (Tables 1 and 2), the proposed minimal data length rule guarantees a good source separation performance with performance indexes (*PI*) under 0.05 in most configurations for FastICA, JADER and AMICA (at least in the plausible case). Extended InfoMax has to be considered separately, as this algorithm requires much more data points. Our decision rule gives minimum data length much smaller than those recommended in literature (over $5n^2$) for high number of channels n, being thus more in adequation with the short time stationarity hypothesis accepted for EEG signals and needed for most ICA algorithms.

A second objective was to evaluate the impact of initialization on the separation performances using whitening or sphering in the first step of these algorithms. Due to the optimization strategy on which they are based, FastICA and JADER show no sensitivity to initialization (decorrelation method). Conversely, natural gradient descent based algorithms AMICA and Extended-Infomax show high sensitivity to initialization. Due to their optimization strategy, these algorithms are much more time consuming and less robust facing outliers, thus requesting an adequate initialization for a reliable convergence with acceptable number of iterations. In the particular case of EEG, modelled as a mixture of dipolar sources, it is possible to initialize the algorithm "near" the solution by sphering. Consequently, the performances of these algorithms improve and they can be reliably applied. In particular, for dipolar mixtures and using sphering as initialization, AMICA showed impressive performances with very low data length even for high number of channels: 3s of data length (512 Hz sampling rate) are sufficient to get an excellent *PI* below 0.03 for the separation of 48 sources. The main drawback of AMICA is its time consumption: it requests more than 60 seconds for this particular example, while FastICA converges in less than 1s to get similar *PI*, although needing 7s of data.

An immediate perspective to this work would be to use more realistic time-structured data, obtained using only modelled neural sources and realistic mixtures (head models). Besides confirming our conclusions for the studied algorithms, this type of simulation setup would allow the evaluation of second order statistics BSS algorithms (SOBI and similar, also widely used for EEG analysis). It might be also useful if algorithms could be tested on more data channels, in order to asses their performances in the context of high-resolution EEG (more than 64 channels).

An interesting perspective, for the specific case of EEG source separation, is to consider new contrasts for BSS algorithms, balancing between dipolarity and independence. Indeed, source independence is known to be often unrealistic for EEGs, as strong synchrony is very likely to appear between distant areas in the brain. A relaxation of the independence constraint might then enhance the EEG source separation performance.

References

1. Schomer, D., Lopes da Silva, F., eds.: Niedermeyers's Electroenephalography: Basic Principles, Clinical Applications and Related Fields. Wolters Kluwer, Lippincott Willimas & Wilkins (2011)
2. Sanei, S., Chambers, J.: EEG Signal Processing. John Wiley & Sons (2007)

3. Scherg, M., Berg, P.: Use of prior knowledge in brain electromagnetic source analysis. Brain Topography **4** (1991) 143–150
4. Cichocki, A., Amari, S.: Adaptive Blind Signal and Image Processing Learning Algorithms and Applications. John Wiley & Sons, New York, USA (2002)
5. Delorme, A., Palmer, J., Onton, J., Oostenveld, R., Makeig, S.: Independent eeg sources are dipolar. PLoS ONE **7**(2) (02 2012) e30135
6. Hyvärinen, A.: Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks **10**(3) (1999) 626–634
7. Bell, A.J., Sejnowski, T.J.: An information-maximization approach to blind separation and blind deconvolution. Neural Computation **7** (1995) 1129–1159
8. Palmer, J., Makeig, S., Delgado, K., Rao, B.: Newton method for the ICA mixture model. In: Acoustics, Speech and Signal Processing, ICASSP 2008. IEEE International Conference on. (31 2008-april 4 2008) 1805 –1808
9. Cardoso, J.: High-order contrasts for independent component analysis. Neural Computation **11**(1) (1999) 157–192
10. Särelä, J., Vigário, R.: Overlearning in marginal distribution-based ICA: analysis and solutions. J. Mach. Learn. Res. **4** (2003) 1447–1469
11. Onton, J., Makeig, S.: Information-based modeling of event-related brain dynamics. Progress in Brain Research **159** (2006) 99–120
12. Delorme, A., Makeig, S.: EEGLab: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of Neuroscience Methods **134**(1) (2004) 9–21
13. Palmer, J., Makeig, S., Delorme, A., Onton, J., Acar, Z.A., Kreutz-Delgado, K., Rao, B.D.: Independent Component Analysis of High-density Scalp EEG Recordings. In: 10th EEGLAB Workshop, Jyväskylä, Finland, June 14-17. (2010)
14. Vaseghi, S., Jetelova, H.: Principal and independent component aanalysis in image processing. (2008)
15. Wu, Y., Wu, B., Liu, J., Lu, H.: Probabilistic tracking on riemannian manifolds. In: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. (dec. 2008) 1 –4
16. Korats, G., Le-Cam, S., Ranta, R.: Impact of window length and decorrelation step on ICA algorithms for EEG blind source separation. In: Biosignals / Biostec INSTICC Annual Conference. (2012)
17. Wendling, F., Bartolomei, F., Bellanger, J.J., Chauvel, P.: Epileptic fast activity can be explained by a model of impaired GABAergic dendritic inhibition. European Journal of Neuroscience **15**(9) (2002) 1499–1508
18. Berg, P., Scherg, M.: A fast method for forward computation of multiple-shell spherical head models. Electroencephalography and Clinical Neurophysiology **90**(1) (1994) 58 – 64
19. Lopes da Silva, F.H., Hoeks, A., Smits, H., Zetterberg, L.H.: Model of brain rhythmic activity. Biological Cybernetics **15** (1974) 27–37
20. Jansen, B., Rit, V.: Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. Biological Cybernetics **73** (1995) 357–366
21. Wendling, F., Hernandez, A., Bellanger, J.J., Chauvel, P., Bartolomei, F.: Interictal to ictal transition in human temporal lobe epilepsy: Insights from a computational model of intracerebral EEG. Clinical Neurophysiology **22**(2) (October 2005) 343–356
22. Hallez, H., Vanrumste, B., Grech, R., Muscat, J., De Clercq, W., Vergult, A., D'Asseler, Y., Camilleri, K., Fabri, S., Van Huffel, S., Lemahieu, I.: Review on solving the forward problem in EEG source analysis. J Neuroeng Rehabil. **4** (2007) 46
23. Ma, J., Gao, D., Ge, F., Amari, S.i.: A one-bit-matching learning algorithm for independent component analysis. In Rosca, J., Erdogmus, D., Príncipe, J., Haykin, S., eds.: Independent Component Analysis and Blind Signal Separation. Volume 3889 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2006) 173–180