Spectral independent component analysis with noise modeling for M/EEG source separation

Pierre Ablin, Jean-François Cardoso, Alexandre Gramfort

Version of Record: https://www.sciencedirect.com/science/article/pii/S0165027021000790
Manuscript_34fb8004a7be1d311cc9ad7f1b9793a8

Spectral independent component analysis with noise modeling for M/EEG source separation☆

Pierre Ablin[a,b], Jean-François Cardoso[c], Alexandre Gramfort[b]

[a] *CNRS and DMA, Ecole Normale Supérieure - PSL University, Paris, France*
[b] *Inria Saclay, Université Paris-Saclay, Palaiseau, France*
[c] *Institut dAstrophysique de Paris, CNRS (UMR7095), Paris, France*

Abstract

Background: Independent Component Analysis (ICA) is a widespread tool for exploration and denoising of electroencephalography (EEG) or magnetoencephalography (MEG) signals. In its most common formulation, ICA assumes that the signal matrix is a noiseless linear mixture of independent sources that are assumed non-Gaussian. A limitation is that it enforces to estimate as many sources as sensors or to rely on a detrimental PCA step.

Methods: We present the Spectral Matching ICA (SMICA) model. Signals are modelled as a linear mixing of independent sources corrupted by additive noise, where sources and the noise are stationary Gaussian time series. Thanks to the Gaussian assumption, the negative log-likelihood has a simple expression as a sum of 'divergences' between the empirical spectral covariance matrices of the signals and those predicted by the model. The model parameters can then be estimated by the expectation-maximization (EM) algorithm.

Results: On phantom MEG datasets with low amplitude dipole sources (20 nAm), SMICA makes a median dipole localization error of 1.5 mm while competing methods make an error $\geq$7 mm. Experiments on EEG datasets show that SMICA identifies a source subspace which contains sources that have less pairwise mutual information, and are better explained by the projection of a single dipole on the scalp. With 10 sources, the number of strongly dipolar

☆Research article

sources (dipolarity $> 90\,\%$) is more than $80\,\%$ for SMICA while competing methods do not exceed $65\,\%$.

Comparison with existing methods: With the noisy model of SMICA, the number of sources to be recovered is controlled by choosing the size of the mixing matrix to be fitted rather than by a preprocessing step of dimension reduction which is required in traditional noise-free ICA methods.

Conclusions: SMICA is a promising alternative to other noiseless ICA models based on non-Gaussian assumptions.

Keywords: ICA, EEG, MEG, source separation

Note: changes from the previous version are written in blue

1. Introduction

Magnetoencephalography and Electroencephalography (M/EEG) are popular non-invasive techniques to record brain activity [27, 41]. They capture respectively the magnetic and electric signals produced by active neurons from the scalp surface or close to it. Each M/EEG sensor captures a combination of the different brain activities. The physics of the mixing is well understood: it is a linear process and can be considered instantaneous.

Independent Component Analysis (ICA) [32] is extensively used in neuroscience for processing M/EEG signals [38]. In its simplest form, it models the observed signals as a linear combination of statistically independent signals called *sources*. Remarkably, ICA can identify these sources 'blindly', that is, without prior knowledge of the underlying physics of the system (except linearity). Applied on EEG signals, it separates meaningful brain signals from artifacts (eye blinks, heartbeats, line noise, muscle, ...) [35], making it an algorithm of choice for artifact rejection [53]. ICA is widely used for the same purpose in MEG studies [40, 54, 34, 17].

Beyond artifact removal, ICA is also used to reveal and study brain activity. In [39], ICA is successfully applied to recover evoked and induced event-related

dynamics in EEG signals. In [24], ICA is used to extract brain sources, on which causal relations are exhibited, uncovering directional coupling. In [51], independent EEG sources are used in a machine learning pipeline, predicting epileptic seizures. ICA is used on MEG signals to identify links between function and structure in the brain in [50]. It can also be used on MEG data, coupled with Hilbert filtering, to uncover resting state networks [10].

Finally, ICA components can be mapped to certain brain areas via *source localization*. Indeed, the individual contribution of each source to each sensor can be represented as a topography on the scalp for EEG or on the helmet for MEG. An equivalent current dipole (ECD) can then be fitted to the topography [49], yielding at the same time an estimate of the source location, and its *dipolarity* (how close it can be explained by a focal activity in the brain modeled with a single dipole).

The hypothesis of *independence* of the sources is at the heart of ICA. However, independence is a statistical property which is difficult to quantify on real data. In neuroscience, the most widely used algorithms are Infomax [7] and FastICA [31]. These algorithms perform *non-Gaussian* ICA: they quantify independence on the marginal (instantaneous) distribution of the data. They ignore any time correlation and focus entirely on the non-Gaussianity of the data. In this case, the sources can be recovered when at most one source has a Gaussian density [15]. Brain sources and artifacts are usually heavy-tail signals which depart form Gaussianity, justifying the use of non-Gaussian algorithms for M/EEG processing.

Another route to ICA is to leverage the time correlation of the sources. In this case, the sources can be recovered when the sources are *spectrally diverse*, that is, when their power spectrum are non-proportional [47]. Among these algorithms, Second Order Blind Identification (SOBI) [8] is one of the most widely used. It jointly diagonalizes a set of time correlation matrices. Another approach closely related to our work consists in the joint diagonalization of spectral covariance matrices [47]. ICA methods based on joint-diagonalization of second order statistics might be less popular than non-Gaussian ICA methods, but have encountered some success in M/EEG processing. Congedo et al. [16] argues that

Pham's approach [47] should be prefered to SOBI for M/EEG preprocessing, one reason being that Pham's approach does not enforce orthogonal constraints. Finally, in order to leverage both non-Gaussianity and spectral diversity, several methods based on short-time Fourier transform (STFT) have been proposed. For instance, Fourier-ICA [33] leverages both non-Gaussianity and spectral diversity with a hybrid method, consisting of the non-Gaussian ICA of concatenated short-time Fourier transforms. Other approaches consist in joint diagonalization of cospectral matrices (covariance matrices of STFT frames) [16] or work in the wavelet domain [46].

While all these algorithms rely on various independence measures, they make the strong assumption that there is no *sensor noise*: they assume that the signal of each sensor is a linear and noiseless combination of sources. A consequence of the noiseless model is that it enforces that there are as many sources as sensors, while the number of sensors is generally fixed by hardware constraints and not by the actual number of brain or artifactual sources present in the data. Unfortunately these noiseless models lead to a degenerate likelihood when there are fewer sources than sensors. This is why, when fewer sources than sensors are expected to be present in the data, a dimension reduction technique like Principal Component Analysis (PCA) is often applied before ICA. However, this two-stage approach, consisting of first applying PCA and then ICA, is heuristic as based on the assumption that independent sources have high variance, which is not necessary. As it is argued in [5] applying PCA before ICA can degrade the quality of the recovered sources. To avoid relying on PCA from dimensionality reduction, it is also sometimes suggested to simply discard some channels. Throwing away data without a clear motivation is arguably questionable.

In order to alleviate this problem, some ICA algorithms incorporate a noise model. As explained in [30], when the noise statistics are known, maximizing the likelihood of such a model is an optimization problem sharing many similarities with dictionary learning [43]. Such procedure is typically much more costly than regular ICA, and it is seldom used in M/EEG processing (See [6, 37] for instance). In [45], noise is modelled in a non-stationary framework: the sources

-and noise- are assumed non-stationary (their instantaneous variance varies over time). The model is estimated by fitting it to the data, via the minimization of a simple quadratic criterion, which deviates from the probabilistic model.

In this article, we study Spectral Matching ICA (SMICA) for M/EEG processing. This ICA model has been first investigated in astronomy for separation of the cosmic microwave background [13, 19]. SMICA models the observations as a sum of a linear mixture of independent sources and noise. It assumes that the sources and noise are Gaussian, and that the sources have non-proportional spectra. This assumption makes it well suited for brain rhythms and artifacts extraction as they are known to have prototypical spectra. Brain sources tend to exhibit so-called "1/f" power spectral densities, while the spectra of artifacts are often localized in certain frequency bands (e.g. muscle artifacts or line noise). Importantly, the statistics of the noise are parameters of the model, and are estimated along the other parameters of the model. Thanks to its noise model, SMICA can estimate fewer sources than sensors without preprocessing for dimension reduction. The sources can be estimated by Wiener filtering, which takes the noise estimation into account and denoises the sources.

The article is organized as follows. In section 2, the SMICA statistical model is introduced and the estimation strategy based on an Expectation-Maximisation (EM) algorithm is described. In section 3, the usefulness of SMICA is demonstrated on various MEG and EEG datasets.

Notation The trace of a matrix $M \in \mathbb{R}^{p\times p}$ is $\mathrm{Tr}(M)$, and its determinant is $|M|$. A matrix is invertible when $|M| \neq 0$, and we write $M \in \mathrm{GL}_p$. Given a vector $u \in \mathbb{R}^p$, the matrix $\mathrm{diag}(u) \in \mathbb{R}^{p\times p}$ is the matrix containing the elements of u on its diagonal, and 0 elsewhere. If M is a $p \times p$ matrix, then $\mathrm{diag}(M)$ is the diagonal matrix with the same diagonal as M. Given $A \in \mathbb{R}^{p\times q}$, the vectorization of A is a vector $\mathrm{vec}(A) \in \mathbb{R}^{pq}$ of entries $\mathrm{vec}(A)_{i+p(j-1)} = A_{ij}$. The Moore-Penrose Pseudo-Inverse of a tall matrix $A \in \mathbb{R}^{p\times q}$ is $A^\dagger = (A^\top A)^{-1}A^\top$.

2. A maximum likelihood approach to noisy ICA

This section introduces our approach to blind source separation for noisy observations (Sec. 2.1). Its application is then discussed in detail (Sec. 2.3).

2.1. The SMICA method in theory

In a noisy ICA model, the outputs of p sensors, *e.g.* M/EEG recordings, collected in a vector $X(t) \in \mathbb{R}^p$, are modelled as noisy instantaneous mixtures of q independent sources represented by a vector $S(t)$ of size q with an additive noise term $N(t)$ of size p, this is

$$X(t) = AS(t) + N(t), \tag{1}$$

where A is the $p \times q$ mixing matrix. The noise is assumed independent from the sources and uncorrelated across sensors.

This model readily translates into the spectral domain. Recall that for a zero-mean p-dimensional stationary time series, $\{X(t)\}$, the $p \times p$ autocovariance matrix $\mathbb{E}[X(t)X(t+\tau)^\top]$ does not depend on t and that its Fourier transform[1]:

$$C(f) = \sum_\tau \mathbb{E}[X(t)X(t+\tau)^\top]\, e^{-2i\pi f\tau} \tag{2}$$

defines $p \times p$ *spectral covariance matrices* $C(f)$. The diagonal entry $C_{aa}(f)$ is the power spectrum of $\{X_a(t)\}$ while $C_{ab}(f)$ contains the cross-spectrum between $\{X_a(t)\}$ and $\{X_b(t)\}$.

The linear relation between data and sources of Eq. (1) translates into the spectral model

$$C(f) = AP(f)A^\top + \Sigma(f) \tag{3}$$

where $P(f)$ and $\Sigma(f)$ are the spectral covariance matrices of sources and of the noise. In this work, we assume that the sources and the noise terms are independent, which means that $P(f)$ and $\Sigma(f)$ are **diagonal** matrices. That property is the key feature enabling the blind estimation of the model since it

[1] For simplicity we have set the sampling period to one time unit.

is how statistical independence between sources is expressed. Incidentally, it also entails that the spectral covariance matrices $C(f)$ are real-valued. Indeed, matrices $P(f)$ and $\Sigma(f)$ are real-valued since they are diagonal matrices and their diagonals are made of power spectra and the mixing matrix A also is real-valued. It follows from Eq. (3) that, unlike generic spectral matrices, a matrix $C(f)$ only has real entries, as a consequence of our statistical model.

This particular structure of the spectral covariance matrices is preserved when spectra are averaged over frequency bands. Define B frequency intervals $I_1, \ldots, I_B$ by $I_b = [f^b_{\min}, f^b_{\max}]$ and consider frequency averages over those bands:

$$C_b = \frac{1}{f^b_{\max} - f^b_{\min}} \int_{f^b_{\min}}^{f^b_{\max}} C_X(f) df \tag{4}$$

Then, upon averaging, Eq. (3) becomes

$$C_b = A P_b A^\top + \Sigma_b \tag{5}$$

where P_b and Σ_b denote the corresponding averages for $P(f)$ and $\Sigma(f)$. As a consequence, the noisy ICA model in Eq. (1) is transformed in the simpler model of Eq. (5), where the parameters are the mixing matrix A, the source powers in each band P_b, and the noise powers in each band, Σ_b.

The noisy ICA model is inferred from by connecting the spectral matrices C_b of model (5) to samples estimates. If T data samples $X(0), \ldots, X(T-1)$ are avialble, spectral matrices are classically estimated from the Fourier coefficients

$$\tilde{\mathbf{x}}_k = \frac{1}{\sqrt{T}} \sum_{t=0}^{T-1} X(t)\, e^{-2i\pi kt/T} \tag{6}$$

by averaging over the relevant frequency bands. These estimates are:

$$\widehat{C}_b = \frac{1}{n_b} \sum_{k:\, \frac{k}{T} \in I_b} \mathrm{Re}\big(\tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^H\big) \tag{7}$$

where $n_b = \#\{k : \frac{k}{T} \in I_b\}$ denotes the number of Fourier coefficients available in band b.

The set $\theta = (A, P_1, \ldots, P_B, \Sigma_1, \ldots, \Sigma_B)$ of all unknown parameters can be estimated by adjusting the model $C_b = A P_b A^\top + \Sigma_b$ to the data as summarized

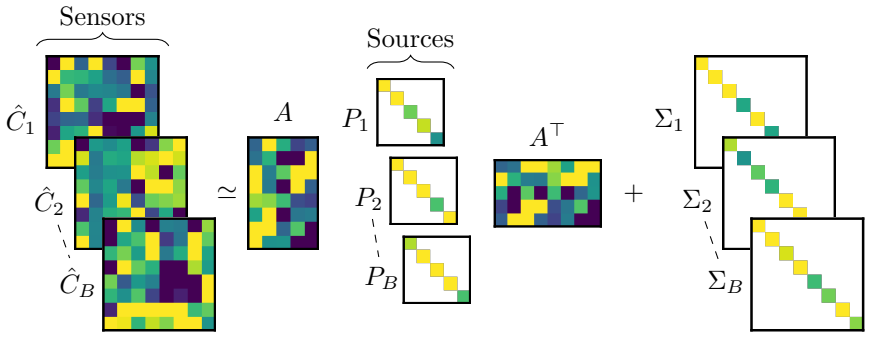


Figure 1: The SMICA method: empirical spectral covariance matrices $\hat{C}_1, \ldots, \hat{C}_B$ computed from the M/EEG data are matched by the model $\hat{C}_b \simeq AP_bA^\top + \Sigma_b$, where the P_b and Σ_b are diagonal positive matrices. Matching is performed in a statistically sound way, by using a matching criterion derived from a simple likelihood.

by $\widehat{C}_b$. This spectral matching principle is illustrated by Figure 1. We advocate using a specific spectral matching criterion:

$$\mathcal{L}(\theta) = \sum_{b=1}^{B} 2n_b \,\mathrm{KL}\left(\widehat{C}_b,\; AP_bA^\top + \Sigma_b\right) \tag{8}$$

where KL is the Kullback-Leibler divergence between two $p \times p$ positive matrices:

$$\mathrm{KL}(C_1, C_2) = \frac{1}{2}\Big(\mathrm{Tr}(C_1C_2^{-1}) - \log\det(C_1C_2^{-1}) - p\Big). \tag{9}$$

The KL-divergence $\mathrm{KL}(C_1, C_2)$ is non-negative and cancels if and only if $C_1 = C_2$.

The particular measure (8) of spectral adjustment between data and model has been chosen because it is (up to an irrelevant constant) asymptotically (for large T and narrow bands) equal to minus the log likelihood of a Gaussian stationary model. Hence, the SMICA estimates inherits some good properties of maximum likelihood estimates, in particular they enjoy a built-in scale invariance and they can be easily computed using the EM algorithm. Those properties and other considerations are discussed in the remaining of this section. The derivation of the SMICA criterion from a likelihood function is explained in Appendix A.

2.2. Statistical properties of SMICA

The novelty in SMICA comes from its noise model: ICA methods that work in the spectral domain are common. In fact, one recovers a popular method by removing the noise in Eq. (5), and by assuming a square mixing matrix $p = q$, that is, finding as many sources as sensors. This was considered by Pham who used the Whittle approximation in his seminal paper [47]. In that case ($p = q$, $N(f) = 0$), the spectral mismatch can be shown to reduce to a joint-diagonality criterion. Estimating A amounts to solving the *joint-diagonalization* of the spectral covariances: A is such that the matrices $A^{-1}\widehat{C}_b A^{-\top}$ are as diagonal as possible. The use of such approach for EEG processing is advocated in [16]. In the experimental section 3, we compare SMICA to this approach of a plain joint diagonalization of spectral matrices and we refer to it as *JDIAG*.

The noise model in SMICA makes the estimation of the model parameters harder than joint-diagonalization, but it enables finer source estimation, through *Wiener filtering* ([11], Chapter 4). In noiseless ICA models of the form $X = AS$, the natural source estimates simply is $\widehat{S} = A^{-1}X$. SMICA can employ the same technique for recovering the source, albeit replacing A^{-1} by $A^{\dagger}$, the Moore-Penrose pseudo-inverse of A. However, the availability of a noise term in SMICA offers a more attractive possibility: one can compute the expected value of the sources given the parameters: $\widehat{S} = \mathbb{E}[S|X, \theta]$ which yields the lowest noise contamination among all unbiased estimators. This is the well known Wiener filtering which has a simple explicit expression in the SMICA model: in each frequency band, it reads

$$\widehat{S}_b = (A^{\top}\Sigma_b^{-1}A + P_b^{-1})^{-1}A^{\top}\Sigma_b^{-1}X_b \ , \tag{10}$$

where X_b is the matrix of signals filtered in the frequency band $[f_b, f_{b+1}]$. This operation is linear in X, and is adaptive to the level of noise: if in frequency band b the estimated noise Σ_b for a sensor is large, then its contribution in the source estimate shrinks towards 0. Note that the standard ICA source estimation formula is recovered when the noise is equal on all channels and tends to 0 (i.e. $\Sigma_b = \lambda_b I_p$ with $\lambda_b \to 0$), yielding at the limit $\widehat{S} = A^{\dagger}X$.

Next, we study the *identifiability* conditions for SMICA. The issue of identifiability is to find the conditions allowing for a unique recovery of the sources. Identifiability conditions of *noiseless* spectral ICA are well established [47]: the sources should not have proportional spectra, that is, for any pair (i, j) of sources the quantities $[P_b]_{ii}/[P_b]_{jj}$ should not be constant with respect to $b = 1, \ldots, B$. The identifiability conditions for the SMICA model (1) are more complicated, and the non-proportionality of the source spectra is only a *necessary* condition. Hence, it is well suited to isolate all kinds of artifacts from brain activity: line noise, muscular activity, heartbeats, eye blinks, since their spectra are usually very different from those of brain activity. Among brain sources, some might have similar power spectrum, such as occipital dominant alpha rhythms and therefore cannot be accurately separated by SMICA.

Finally, we discuss the algorithm used to estimate the parameters, that is the specific numerical procedure to minimize the spectral matching criterion (or maximize the likelihood function) of Eq. (8). Since this is a likelihood function with latent (or unobserved) variables —namely the source and noise signals— it can be optimized by resorting to the celebrated Expectation-Maximization (EM) algorithm [21]. The EM algorithm is appealing because it does not require any hyperparameters like learning rates, and is guaranteed to decrease the loss function at each iteration. Still, this approach is generally slow (it might require many iterations to reach a satisfactory set of parameters), and other optimization techniques could be investigated for the fast minimization of the negative log-likelihood $\mathcal{L}$. The EM algorithm for SMICA is described in the appendix (section Appendix C).

Next, we discuss the practical application of SMICA for M/EEG processing.

2.3. SMICA for M/EEG processing

This section goes through various considerations regarding the application of SMICA to M/EEG signals. A first advantage of SMICA over other ICA algorithms for M/EEG processing is its embedded dimension reduction. It is often the case that there are more sensors than sources which can be significantly

recovered from the observations. When $p > q$ matrix A cannot be inverted and one usually performs some dimension reduction in order to apply ICA algorithms which require a square matrix ($p = q$) to operate. In SMICA, thanks to the presence of a noise term, spectral covariance matrices are not degenerate even if $q < p$ and a non-square matrix A can be fitted by SMICA in a statistically sound way even when there are fewer sources than sensors, without resorting to a pre-processing stage of dimension reduction. In some sense, one can say that SMICA has a *built-in dimension reduction* of the signal part because it can fit a tall ($q < p$) mixing matrix. It often argued that reduction dimension in EEG processing deteriorates the quality of the subsequent ICA decomposition [5]. SMICA offers a simple way to circumvent this problem, by embedding dimension reduction in the ICA in a statistically sound way.

Signal denoising using SMICA. Like any ICA algorithm, SMICA estimates sources which can be marked as spurious / non-biological by specialists: in addition to brain sources, it usually recovers physiological artifacts (heartbeats, eye blinks) and external electromagnetic perturbations (room and line noise). The remaining sources can then be projected back in the signal space, giving clean M/EEG signals. Thanks to its noise modeling, SMICA makes these two operations statistically sound: the sources $\widehat{S}$ are estimated by Wiener filtering, some sources can be manually or automatically marked as spurious. If the source i is marked as spurious, we set $\widehat{S}_i = 0$. The cleaned M/EEG signals are computed as $X_{cleaned} = A\widehat{S}$.

Combining SMICA with another ICA algorithm. In practice, it might happen that some sources recovered by SMICA have similar spectra, which indicates that these sources are not well separated. Another ICA procedure (typically, based on non-Gaussianity) can then be applied on these sources in order to better separate them.

Taking it a step further, SMICA can also be used as a *source subspace* identifier. Applied on the M/EEG signals $X \in \mathbb{R}^{p \times T}$, SMICA produces a mixing matrix $A_1 \in \mathbb{R}^{p \times q}$ and a source matrix $S_1 \in \mathbb{R}^{q \times T}$ estimated by Wiener filtering.

The sources in S_1 are maximally independent with respect to the separation criterion of SMICA, but might not be maximally independent with respect to another criterion. Applying another ICA algorithm on S_1 yields a square mixing matrix $A_2 \in \mathbb{R}^{q \times q}$, and a new source matrix $S_2 = A_2^{-1} S_1$. The overall mixing matrix linking the sources S_2 and the original dataset X is the matrix product $A_{Total} = A_1 A_2$. For instance, using a non-Gaussian ICA algorithm like Infomax or FastICA on the sources found by SMICA may disentangle sources that share similar power spectrum. The practical benefits of such approach are demonstrated in the experiments presented in section 3.

Complexity of SMICA. Estimation of the parameters of SMICA is a two step process: first, the empirical spectral covariances $\hat{C}_1, \ldots, \hat{C}_B$ are computed, and then the EM algorithm is used to infer the parameters, based solely on those covariances. Only the first step scales with the length of the signal T: the second only depends on the dimensions of the problem p, d and the number of bins B. In practice, we find that in the setting of our experiments, computing the covariances takes a negligible time compared to the time it takes to infer the parameters with the EM algorithm. The complexity of the EM algorithm does not depend on the number of samples T; only the computation of the covariances does. This differs from non-Gaussian ICA algorithms, for which the estimation time is roughly proportional to the length of the recordings.

In practice, we found that fitting SMICA on a 102 sensors MEG dataset with 40 frequency bins and 100 sources takes about 15 minutes using one CPU of a recent laptop.

Frequency selection. SMICA exploits spectral information, but not necesarily over the whole frequency range. Typically, one may exclude the highest frequencies if they are dominated by noise or cut off by a sampling filter. One may also ignore the very lowest frequencies if they are dominated by slow drift artifacts. In general, there is no counter indication to restricting the sum (8) to the high SNR part of the frequency range.

3. Experiments

We report some experiments comparing our approach to other (noiseless) ICA algorithms: these algorithms all model the dataset X as $X = AS$, where $A \in \mathbb{R}^{p \times p}$ is the *square* mixing matrix, and S is the source matrix. However, they estimate the independent components based on different independence criterion.

We start by briefly describing the ICA algorithms which are compared to SMICA.

Non-Gaussian ICA. Non-Gaussian ICA algorithms model the source time-series as independent and identically distributed, with non-Gaussian probability density functions. Some of the most popular ICA algorithms fall in this category: FastICA [31], Infomax [7], its extended version [36], JADE [14] and more recently AMICA [44].

Second order blind identification. The SOBI algorithm [8] aims at recovering sources with spectral diversity, just like SMICA and JDIAG. It does so in a heuristic way, by joint-diagonalization of a set of correlation matrices $\frac{1}{T-\tau}\sum_{t=1}^{T-\tau} X(t)X(t-\tau)^\top$ for a set of time lags $\tau_1, \ldots, \tau_B$, rather than spectral covariance matrices. Choosing an appropriate set of time lags is not an obvious task; we use the set advised in [52]. Unlike JDIAG, the joint diagonalization criterion is ad-hoc, and does not correspond to a principled statistical criterion. For instance, since JDIAG follows the maximum-likelihood principle, it is asymptotically Fisher-efficient and reaches the Cramer-Rao lower bound, unlike SOBI. This is why several articles argue for the use of JDIAG rather than SOBI [22, 16].

Estimating fewer sources than sensors using PCA. Contrarily to SMICA, algorithms described above can only estimate as many sources as sensors. Therefore, in order to estimate fewer sources, a dimension reduction step should be performed beforehand. Principal Component Analysis is the algorithm of choice for this task. The components are chosen to explain as much variance in the data as possible. Since this method is blind to higher order interactions, it might discard

some sources which are important but of low power. As a consequence, Artoni et al. [5] argue that applying PCA before ICA leads to degraded decomposition.

Numerical setup. In our experiments, we take Infomax as the reference non-Gaussian ICA algorithm with the fast and robust optimization algorithm Picard [2, 1] and using $\tanh(\cdot)$ as the non-linear activation function. The joint-diagonalization algorithm for SOBI is a combination of [55] with a backtracking line-search. The joint-diagonalization algorithm for JDIAG uses the fast implementation described in [3]. In all experiments we set the frequency bins F_b of SMICA and JDIAG as uniform in the range $1 - 70$ Hz, with 40 bins. The M/EEG analysis is conducted using the Python package MNE [26, 25]. Figures are made using Matplotlib [29].

The python code for SMICA is available online at `https://github.com/pierreablin/smica`.

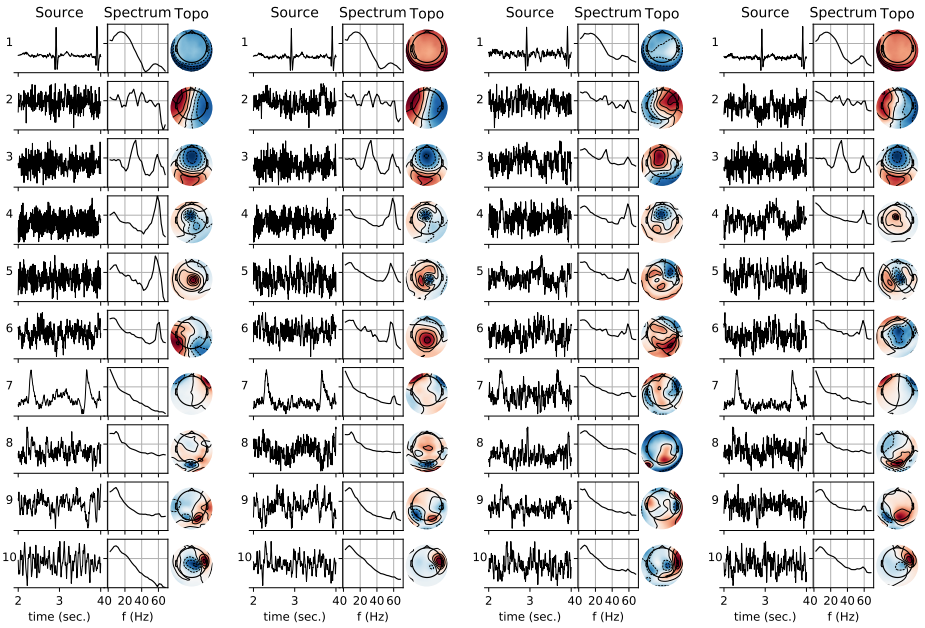
3.1. Qualitative comparisons

3.1.1. Comparisons on a MEG dataset

We start by showing the decomposition found by SMICA, JDIAG, SOBI and Infomax on a MEG dataset, where the subject was presented checkerboard patterns into the left and right visual fields, as well as monaural auditory tones to the left or right ears. Stimuli occurred every 750 ms (See Gramfort et al. [26] for a description of the dataset). MEG is acquired with 102 magnetometers and 204 gradiometers.

For this experiment, we only consider the 102 magnetometer channels. Each ICA algorithm returns 40 sources (after PCA for JDIAG, SOBI and Infomax). We hand pick 10 sources to display in Figure 2, which shows their time-course, power spectrum and topography.

SMICA isolates heartbeats (source 1), eye blinks (source 7) and brain (sources 8-10) from line noise: only source 9 shows a very small peak at 60 Hz. Source 7 and 9 in JDIAG's decomposition, source 1, 7, 9, 10 for SOBI and source 1, 8, 9 for Infomax do show a higher peak at 60 Hz suggesting that line noise leaks



(a) SMICA (b) JDIAG (c) SOBI (d) Infomax

Figure 2: Different ICA decompositions on MEG data. Source 1 corresponds to heartbeats, sources 2-6 to environmental noise, with strong peaks around 60 Hz. Source 7 corresponds to eye blinks. Sources 8, 9 correspond to occipital alpha rhythm (excect for SOBI which did not find such sources). Source 10 corresponds to a dipolar auditory source. Sources 7, 9 found by JDIAG present artifactual 60 Hz components, like source 1, 9 for Infomax. SOBI does not properly identify the eye-blink source.

into these others sources. SOBI fails to separate properly the eye-blinks, and gives decompositions that differ substantially from other algorithms. This first experiment demonstrates that SMICA does reveal expected artifactual sources, both physiological and environmental, and that they potentially leak less into the valuable neural ones. More quantitative evidence is provided below.

3.1.2. Comparisons on an EEG dataset

We run SMICA and JDIAG on a 69-channel EEG data coming from the dataset described in [20]. Both algorithms return 20 sources, which are displayed in Figure 3. Differences between SMICA and JDIAG are now more striking,

probably due to the greater noise level compared to the MEG recording.

Although decompositions differ in many aspects, we want to focus on source 2 recovered by SMICA, which is not found by JDIAG. There is a sharp peak at 60 Hz, indicating that it likely corresponds to line noise. The second peak at 50 Hz may seem spurious but is most probably due to spectral aliasing. Indeed the fifth harmonic of a line at 60 Hz when sampled at 250 Hz appears at frequency $5 * 60 - 250 = 50\,\text{Hz}$. To test whether it is a plausible line source, we resort to a separate experiment: we determine the spatial filter $w \in \mathbb{R}^{59}$ such that the time series wX is of power 1 with a maximal power at 60 Hz. This method is called Spatio-Spectral Decomposition (SSD) [42]. To do so, we filter X in a narrow band around 60 Hz, yielding signals X_f. Then, we find w by maximizing the power of wX_f under the constraint that the power wX is 1 which is done by maximizing the Rayleigh quotient between the covariance of X_f and of X.

The power spectrum of the corresponding source wX is displayed in Figure 3, along with the power spectrum of the source number 2 found by SMICA. The 50 Hz aliased harmonic is also recovered by SSD, suggesting that the source recovered by SMICA isolates the line signal correctly.

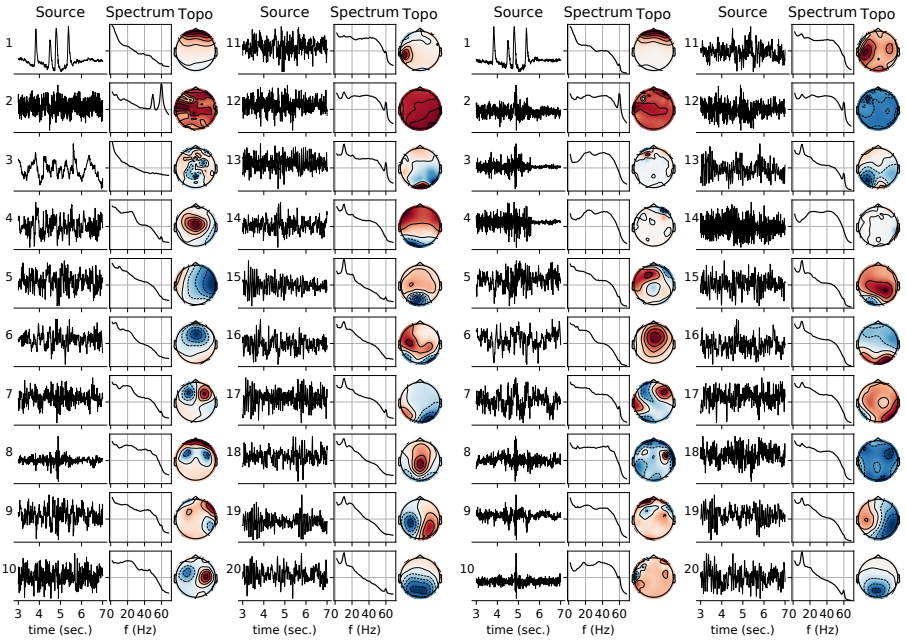
3.2. Quantitative results on large datasets

3.2.1. Experiment on MEG Phantom data

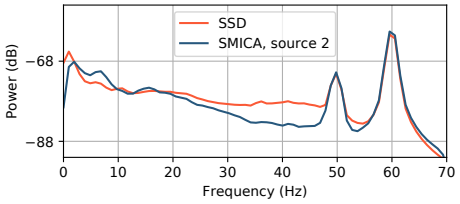
In this section, we experiment with MEG Phantom data. The recording comes from a fake plastic head with electric dipoles. Dipoles emit sinusoidal pulses at a fixed frequency 20 Hz for 0.5 second (10 periods) and are then turned off for 1 second. This is repeated for 150 seconds.

We have 24 datasets. Each dataset corresponds to one dipole location among 8 and one amplitude (either 1000, 200 or 20 nAm). The amplitude corresponds to the peak-to-peak difference. We cut each dataset in half to obtain twice as many datasets.

As the true locations of the sources in the phantom are known one can map the origin of an ICA source by fitting an equivalent current dipole (ECD) to the source topographies and evaluate the localization errors. On each dataset, we



(a) SMICA (b) JDIAG



(c) Spectrum of source 2 of SMICA and source obtained by Spatio-Spectral Decomposition (SSD) at 60Hz.

Figure 3: Comparison of SMICA and JDIAG on an EEG dataset. Both algorithms return 20 sources. The ordering of the sources of SMICA is made by hand; the ordering of the sources of JDIAG is made by maximizing the correlation with SMICA's sources. Both algorithms accurately recover the eye blinks (source 1). For SMICA, sources 4-20 correspond to brain activity. SMICA finds two dipolar beta-rhythm sources (sources 7 and 10). Each algorithm recovers a source corresponding to line noise, with a large peak at 60 Hz (source 2). For SMICA, there is an additional peak at 50 Hz, which is not an artifact. Spatio-Spectral Decomposition (SSD) tuned to the source of maximal power at 60 Hz yields a similar peak at 50 Hz (c).

apply SMICA, Infomax and JDIAG to obtain 40 sources. For each source, we fit an ECD, and only keep the source corresponding to the closest location to the true dipole.

Besides ICA, Maxwell filtering can also be used for dipole localization. After Maxwell filtering, we compute the evoked potential and fit a dipole at the peak.

Finally, we employ Spatio-Spectral Decomposition (SSD), as described in section 3, to find a linear combination of sensors with maximal amplitude at 20 Hz. This method incorporates more knowledge of the problem than others, because we provide it with the dipole frequency.

In Figure 4, we display the average distance to the true dipole and the residual variance in the dipole fit, for each dipole amplitude (1000, 200 and 20 nAm). The localization errors increase as the sources amplitude diminishes for each algorithm. Yet, for the smallest amplitude (20 nAm), which is the most challenging, SMICA outperforms all other methods in terms of localization (note the logarithmic scale).

3.3. SMICA finds highly dipolar and independent source subspaces

In this section, we illustrate the ability of SMICA to capture meaningful brain sources. We use the same datasets as in [20]. It contains the EEG recording of 15 subjects, with 69 EEG channels. For a target number of independent sources, different ICA procedures described in the article are applied to the datasets. The Wiener and Pseudo-Inverse methods correspond to the combination of SMICA with Infomax, as described in the paragraph 'Combining SMICA with another ICA algorithm' of section 2.3. Wiener corresponds to computing SMICA's sources with Wiener filtering, Pseudo-Inverse corresponds to computing SMICA's sources with pseudo-inversion of the mixing matrix A, as described in the paragraph 'Source estimation by Wiener filtering' of Sec. 2.3. For each decomposition, we compute the dipolarity of each source, as well as the pairwise mutual-information between each pair of sources.

Results for the dipolarity are displayed in Figure 5, results for the pairwise mutual information are displayed in Figure 6.

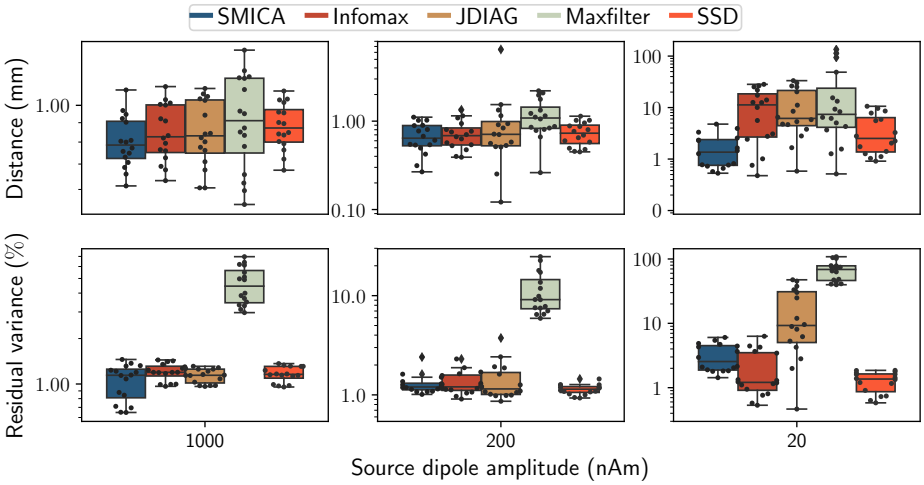


Figure 4: Dipole localization on MEG Phantom dataset. **Top**: Distance between the true dipole and the estimated dipole. **Bottom**: Residual variance in the dipole fit. Each column corresponds to a different source amplitude (from left to right, 1000, 200 and 20 nAm). Dipole fitting is applied on 24 datasets. Each black dot corresponds to a dataset. When the source dipole amplitude is low (20 nAm), SMICA has better localization performance than the other methods.

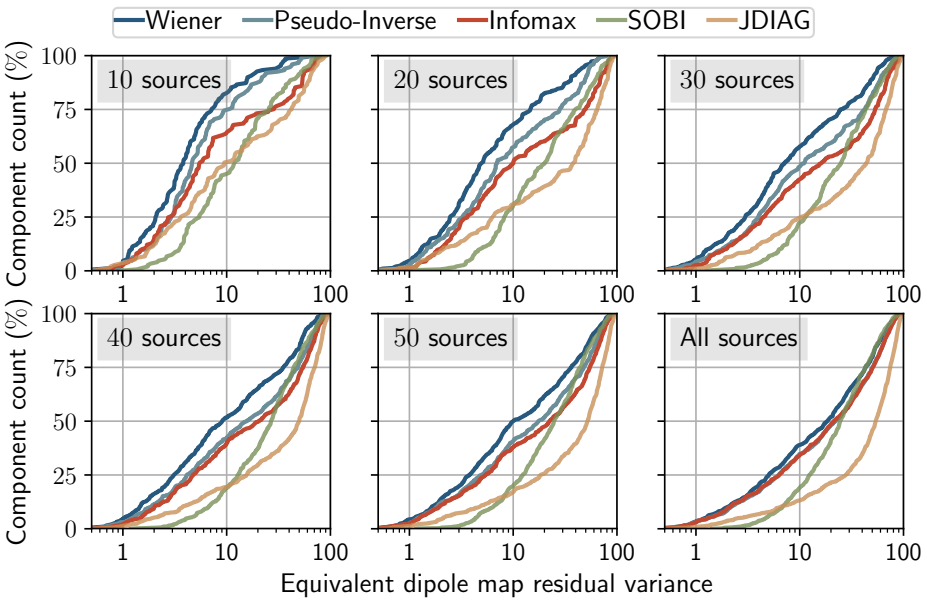


Figure 5: Distribution of equivalent dipole map residual variance for each source returned by each algorithm on the 15 datasets of 69 sensors, where each algorithm returns 10, 20, 30, 40, 50 or 69 sources. For Infomax, SOBI and JDIAG, PCA is first applied on the data matrix to obtain the desired number of channels. Wiener and Pseudo-Inverse correspond to applying Infomax on the sources recovered by SMICA, either by the Wiener or Pseudo-Inverse method. The figure should be understood in the following way. Looking at the first plot, corresponding to 10 sources, we see that about 80% of components found by the method 'Wiener' have an equivalent dipole map residual variance lower than 10%. About 65% of components found by 'Infomax' have an equivalent dipole map residual variance lower than 10%. Overall, the method 'Wiener' finds more dipolar components, for every number of sources considered.

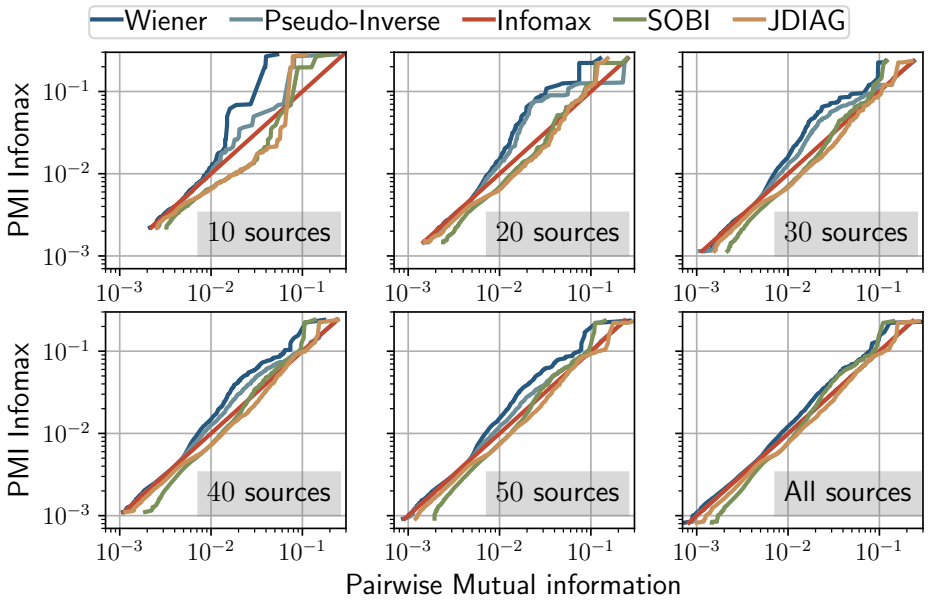


Figure 6: Pairwise mutual information (PMI) for different ICA decompositions returned by each algorithm on the 15 datasets of 69 sensors, where each algorithm returns 10, 20, 30, 40, 50 or 69 sources. For Infomax, SOBI and JDIAG, PCA is first applied on the data matrix to obtain the desired number of channels. Wiener and Pseudo-Inverse correspond to applying Infomax on the sources recovered by SMICA, either by the Wiener or Pseudo-Inverse method. PMI is displayed in a Q-Q plot showing the quantiles of the distribution of PMI found by each algorithm against the quantiles of distribution of PMI found by Infomax. Infomax therefore corresponds to the line $x = y$. Algorithms above the line $x = y$ have less PMI remaining than Infomax, and hence obtain more independent sources than Infomax as quantified by PMI. Here again the the Wiener approach is particularly competitive, especially when few sources are estimated.

The results of raw SMICA are not displayed here, as it recovered less dipolar sources than Infomax. That is not unexpected if one follows the discussion of [20], arguing that maximum-likelihood non-Gaussian methods, like Infomax, are the ICA methods that recover the most dipolar sources.

For this experiment, using SMICA as a subspace identifier to perform dimension reduction yields the best results, regardless of the number of sources that are recovered. The benefits of the Wiener filtering over pseudo-inversion are also illustrated: even without dimension reduction, it denoises the signals, which leads to improved ICA decompositions.

4. Discussion

We now discuss the advantages and shortcomings of SMICA for M/EEG processing.

Noise model. Results in Sec. 3.1.1 (Figure 2) have shown that SMICA can better isolate stationary artifacts (environmental line noise, muscule activity, eye blinks, heartbeats,...) than non-Gaussian ICA methods. This can be explained by the fact that artifacts and brain sources usually have very different spectra. Artifacts tends to have peaks or bumps in the spectra in higher frequencies while neural sources have spectra with power laws or exponential decays [12, 18, 48]. The rate of decay of the different spectrum of neural sources also depends on the underlying source. This spectral diversity can be exploited by SMICA to improve source separation. Overall, this method is well suited to separate artifacts from neural activity. We would also like to stress that these artifacts are captured as *sources* by SMICA: they correspond to "S" in the equation $X = AS + N$, and not to the modelling noise "N". For instance, if an external signal, independent from other sources, has a correlated spectral power over the sensors, then SMICA will capture it as a source. By assumption, the modelled noise N only identifies *sensor* noise: noise that is decorrelated across sensors. Because we estimate the spectral power of noise, the model is flexible enough to capture any such signal.

SMICA as a dimension reduction tool. Thanks to the noise model, we have a principled way to perform dimension reduction and to recover the source time courses. In [5], it is argued that PCA is suboptimal for EEG data, and that even channel subsampling is to be preferred. By contrast, the subspace identified by SMICA contains sources that are estimated by incorporating spectral information of the original signals, and these sources have spectral diversity. This benefit is observed in practice. In Sec. 3.3 (Figure 5 and 6) we show that SMICA identifies a source subspace that contains more dipolar and independent components than PCA. As such, SMICA can be a useful tool for dimension reduction.

Comparison with noiseless model. Importantly, as illustrated by Figure 2, SMICA is similar to JDIAG for clean data, since the noise subspace in this case is simple to find with a PCA. In most practical applications, like clinical recording processing, signals are usually contaminated with strong sensor noise, with a $1/f$ spectral signature. In frequency bands where noise is strong, the spectral covariances $\hat{C}_b$ have large diagonal coefficients, which artificially skews noiseless joint-diagonalization methods towards diagonal mixing matrices. This problem is discussed at length in Congedo et al. [16, Appendix B], in which it is proposed to weight the covariances matrices in the joint-diagonalization criterion with a quantity that measures a deviation to diagonality: using this trick, the covariances $\hat{C}_b$ that have a strong diagonal are less taken into account in the criterion. SMICA provides an alternative to this ad-hoc method, by having a statistically sound noise model. The strong diagonality of covariance matrices is automatically taken into account by SMICA, by estimating a high noise power Σ_b.

Other spectral approaches. This work is not the first one using spectral ICA and highlighting its benefits for M/EEG recordings. Some works have focused on convolutional mixtures (for which using the Fourier domain turns convolutions into products). In [23], authors use convolutive ICA in time-domain, while [4] use a complex Infomax to find travelling waves. In the end, these methods estimate one mixing matrix for each frequency bin. However, independent models in each band might fail to recover brain rhythms with several frequency peaks. For

instance, mu-rhythm is characterized by concurrent activity near 10 Hz and 20 Hz [28, 41].

Automatic estimation of the number of sources?. Some improvements to the current SMICA algorithm can be investigated. First, since it comes from a principled statistical framework, it would be interesting to implement a data-driven way of computing the most likely number of sources in the data: an algorithm to automatically select the correct number of sources. However, preliminary experiments show that usual statistical criteria like Akaike Information Criterion or Bayesian Information Criterion are not satisfactory in this setting, likely because the model is not complex enough to explain fully M/EEG signals. The EM algorithm for fitting SMICA is also quite slow, some improvements could be possible by further studying the geometry of the cost function and proposing quasi-Newton algorithms, as done recently for Infomax [2].

Possible extensions of SMICA. SMICA could be extended in several interesting ways. In MEG acquisition, the empty room is sometimes recorded before the experiment. In this case, we could assume that the noise term in Eq. (1) shares the same spectral signature as the empty room recording: the matrices Σ_b are no longer estimated by the model, but are now taken as the spectral covariance matrices of the empty room. Only the mixing matrix A and the source powers P_b are left to estimate. The sources estimated by the algorithm should then only correspond to biological sources; in particular they should automatically be cleaned from line noise.

In order to obtain a pure subspace identification method, one could also drop the hypothesis of independence between sources and minimize the spectral matching criterion with respect to any positive spectral covariance matrices P_b, instead of constraining them to be diagonal.[2].

Finally, the proposed method could also be extended to non-stationary signals,

[2]In this setting, it is possible, but not mandatory, to constrain matrix A to have orthonormal columns: $(A^\top A = I_q)$.

where the spectral covariance matrices are replaced by time-lagged covariance matrices. The resulting algorithm would resemble the procedure proposed in [45], but with a proper likelihood-based estimation rather than the ad-hoc criterion proposed by the authors. More generally, the spectral model with noise and EM algorithm can be employed to recover the parameters of a noisy ICA model $X = AS + N$ using any kind of second-order statistics.

5. Conclusion

In this work, we have introduced a novel ICA algorithm, SMICA, which adds a noise model to the standard ICA model. By assuming a model of Gaussian stationary sources, we obtained a tractable closed-form likelihood, which can then be maximized with the expectation-maximisation algorithm. Blind identifiability stems from spectral diversity: this likelihood permits the separation of sources with different power spectra. The model can estimate fewer sources than sensors, and the sources can then be recovered by Wiener filtering, which takes noise into account. We then demonstrated the promises of our method for M/EEG signal processing: compared to noise-free ICA algorithms, SMICA extracts sources that are more dipolar and have less pairwise mutual information. Besides when applied on controlled recordings using an MEG phantom, SMICA is able to locate dipoles more precisely, especially for low amplitude sources. These results indicate that SMICA is a promising alternative to other more standard ICA algorithms for M/EEG signal denoising and exploration.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation program (Grant agreement SLAB ERC-StG-676943). P.A. acknowledges support from the French government under management of Agence Nationale de la Recherche as part of the Investissements d'avenir program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute).

Competing interests statement

This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation program (Grant agreement SLAB ERC-StG-676943). P.A. acknowledges support from the French government under management of Agence Nationale de la Recherche as part of the Investissements davenir program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute).

Credits

Pierre Ablin: Conceptualization, methodology, software, investigation, writing - original draft, visualization.

Jean-François Cardoso: Conceptualization, methodology, software, investigation, writing - review and editing, supervision.

Alexandre Gramfort: Conceptualization, methodology, software, writing - review and editing, supervision, funding acquisition.

Appendix A. Spectral mismatch and likelihood

We show the statistical origin of the measure (8) of spectral mismatch. The starting point is that, for a zero-mean p-variate stationary times series with spectral covariance matrix $C(f)$, the Fourier coefficients $\tilde{\mathbf{x}}_k$ defined at Eq.(6) have zero mean and covariance $\mathbb{E}[\tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^H] = C(k/T)$. Next, asymptotically (for large T), these coefficients also are normally distributed and pairwise uncorrelated [9]. The *Whittle approximation* to the likelihood consists in assuming that these properties hold even in finite sample size. In that case, the probability density for the Fourier coefficients $(\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_{T/2})$ is the product over frequencies of Gaussian densities: $\log p(\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_{T/2}) = \sum_{k=1}^{T/2} \log \mathcal{CN}(\tilde{\mathbf{x}}_k; C(k/T))$ where $\mathcal{CN}(\mathbf{x}; C)$ denotes the complex circular Gaussian density for a zero-mean vector $\mathbf{x}$ of covariance matrix $C = \mathbb{E}[\mathbf{x}\mathbf{x}^H]$. In Appendix B, we give a simple expression (Eq. (B.1)) for

$\mathcal{CN}(\mathbf{x}; C)$ in the special case where $C(f)$ is real-valued, yielding:

$$\log p(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_{T/2}) = -\sum_{k=1}^{T/2} \left(\mathrm{Tr}\big(C(\frac{k}{T})^{-1}(\tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^H)\big) + \log\det(\pi C(\frac{k}{T})) \right) \quad \text{(A.1)}$$

It is worth noting that, in general, the covariance matrix of a circular complex Gaussian vector is complex-valued. However, the noisy stationary ICA model considered in this paper does enjoy real-valued spectral covariance matrices, as seen from Eq. (3).

The final step is to approximate the spectra as constant over thin spectral bands, that is, $C(f) = C_b$ when $f \in I_b$. Then, the sum (A.1) over frequencies can be expressed as a sum over spectral bands:

$$\log p(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_{T/2}) = -\sum_{b=1}^{B} n_b \left(\mathrm{Tr}(\widehat{C}_b C_b^{-1}) + \log\det(\pi C_b) \right).$$

Hence, with the definition (9) of the KL divergence, we obtain

$$\log p(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_{T/2}) = -\sum_{b=1}^{B} 2n_b \,\mathrm{KL}\left(\hat{C}_b, C_b(\theta)\right) + \mathrm{cst} \quad \text{(A.2)}$$

which shows that, up to a constant term, the spectral matching criterion is minus the logarithm of the Whittle likelihood (A.2).

The presence of the factor 2 in (A.2) can be traced back to the fact that a Fourier component at frequency k has 2 degrees of freedom: its real and its imaginary parts, or equivalently, that in real space any Fourier component has a sine and a cosine terms.

Appendix B. Likelihood and complex vectors

We give joint pdf for the real and imaginary parts of a zero-mean complex random $p \times 1$ vector $\mathbf{x} = \mathbf{a} + i\mathbf{b}$ with a *real* covariance matrix:

$$\mathbf{x} = \mathbf{a} + i\mathbf{b} \qquad C = \mathbb{E}[\mathbf{x}\mathbf{x}^H] \in \mathbb{R}^{p \times p}.$$

and whose distribution is invariant under any phase change, that is, $\mathbf{x}$ has the same distribution as $e^{i\phi}\mathbf{x}$ for any angle ϕ. We look at the consequence of this

invariance on the structure of C and give the joint pdf of a and b when they are jointly Gaussian. We start with:

$$C = \mathbb{E}[\mathbf{x}\mathbf{x}^H] = \mathbb{E}[(\mathbf{a}+i\mathbf{b})(\mathbf{a}+i\mathbf{b})^H] = \mathbb{E}[(\mathbf{a}\mathbf{a}^\top + \mathbf{b}\mathbf{b}^\top) \,-\, i\,\mathbb{E}[\mathbf{a}\mathbf{b}^\top - \mathbf{b}\mathbf{a}^\top],$$
$$\text{and}\;\; \mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbb{E}[(\mathbf{a}+i\mathbf{b})(\mathbf{a}+i\mathbf{b})^\top] = \mathbb{E}[(\mathbf{a}\mathbf{a}^\top - \mathbf{b}\mathbf{b}^\top) \,+\, i\,\mathbb{E}[\mathbf{a}\mathbf{b}^\top + \mathbf{b}\mathbf{a}^\top].$$

If $\mathbf{x}$ is changed into $e^{i\phi}\mathbf{x}$, then matrix $\mathbb{E}[\mathbf{x}\mathbf{x}^H]$ is unchanged but $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ is changed into $e^{2i\phi}\mathbb{E}[\mathbf{x}\mathbf{x}^\top]$. However, by the phase invariance, there should be no change. That is only possible if $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = 0$. Therefore $\mathbb{E}[\mathbf{a}\mathbf{a}^\top - \mathbf{b}\mathbf{b}^\top] = 0$ and $\mathbb{E}[\mathbf{a}\mathbf{b}^\top + \mathbf{b}\mathbf{a}^\top] = 0$. Combining that with the assumption of a real covariance matrix $C = \mathbb{E}[\mathbf{x}\mathbf{x}^H]$ which implies $\mathbb{E}[\mathbf{a}\mathbf{b}^\top - \mathbf{b}\mathbf{a}^\top] = 0$ yields

$$\mathbb{E}[\mathbf{a}\mathbf{a}^\top] = \mathbb{E}[\mathbf{b}\mathbf{b}^\top] = C/2 \qquad \mathbb{E}[\mathbf{a}\mathbf{b}^\top] = \mathbb{E}[\mathbf{b}\mathbf{a}^\top] = 0$$

Therefore the joint $(2p) \times (2p)$ covariance matrix for the pair $(\mathbf{a}, \mathbf{b})$ is

$$\mathrm{Cov}\Big(\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix}\Big) = \begin{bmatrix}\mathbb{E}[\mathbf{a}\mathbf{a}^\top] & \mathbb{E}[\mathbf{a}\mathbf{b}^\top]\\ \mathbb{E}[\mathbf{b}\mathbf{a}^\top] & \mathbb{E}[\mathbf{b}\mathbf{b}^\top]\end{bmatrix} = \begin{bmatrix}C/2 & 0\\ 0 & C/2\end{bmatrix}$$

If $\mathbf{a}$ and $\mathbf{b}$ are jointly Gaussian, their probability density $p(\mathbf{a}, \mathbf{b})$ is given by

$$\log p(\mathbf{a},\mathbf{b}) = \frac{1}{2}\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix}^\top \Sigma^{-1}\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix} - \frac{1}{2}\log\det(2\pi\Sigma), \quad \Sigma = \mathrm{Cov}\Big(\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix}\Big) = \begin{bmatrix}C/2 & 0\\ 0 & C/2\end{bmatrix}.$$

The block structure of Σ yields $\log\det(2\pi\Sigma) = 2\log\det(\pi C)$ and also

$$\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix}^\top \begin{bmatrix}C/2 & 0\\ 0 & C/2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix} = 2\,\mathrm{Tr}\big(C^{-1}(\mathbf{a}\mathbf{a}^\top + \mathbf{b}\mathbf{b}^\top)\big).$$

Since C is a real (by assumption) symmetric matrix, we have

$$\mathrm{Tr}\big(C^{-1}(\mathbf{a}\mathbf{a}^\top + \mathbf{b}\mathbf{b}^\top)\big) = \mathrm{Tr}\big(C^{-1}\mathbf{x}\mathbf{x}^H\big) = \mathrm{Tr}\big(C^{-1}\,\mathrm{Re}(\mathbf{x}\mathbf{x}^H)\big)$$

Combining the previous two results yields the joint density:

$$\log p(\mathbf{a},\mathbf{b}) = -\,\mathrm{Tr}\big(C^{-1}\,\mathrm{Re}(\mathbf{x}\mathbf{x}^H)\big) \;-\; \log\det(\pi C) \tag{B.1}$$

Appendix C. The EM algorithm for SMICA

The following computations closely follow those of [19, Appendix C]. The parameters are $\theta = \{A, \Sigma_1, \dots, \Sigma_B, P_1, \dots, P_B\}$, and the latent variables are the sources in each frequency bands $S_1, \dots, S_B$.

E-step. At the E-step, the sufficient statistics of the model are computed. Since the model is Gaussian, they are simply the second-order statistics: $\mathbb{E}[S_b S_b^\top|\theta], \mathbb{E}[S_b X_b^\top|\theta]$ and $\mathbb{E}[X_b X_b^\top|\theta]$. In the following, let $\Gamma_b = (A^\top \Sigma_b^{-1} A + P_b^{-1})^{-1}$ and $W_b = \Gamma_b A^\top \Sigma_b^{-1}$ the Wiener filter. We have:

$$R_b^{XX} \triangleq \mathbb{E}[X_b X_b^\top|\theta] = \hat{C}_b \tag{C.1}$$

$$R_b^{SX} \triangleq \mathbb{E}[S_b X_b^\top|\theta] = W_b \hat{C}_b \tag{C.2}$$

$$R_b^{SS} \triangleq \mathbb{E}[S_b S_b^\top|\theta] = W_b \hat{C}_b W_b^\top + \Gamma_b \tag{C.3}$$

M-step. At the M-step, the parameters of the model θ should be modified in order to decrease the loss function, using the sufficient statistics obtained in the E-step. To do so, we compute the EM functional, $\Phi(A, (P_b), (\Sigma_b), S) = -\log(p(X, S|\theta))$. To begin, we assume that there is only on frequency band ($B = 1$), which gives a Gaussian "white" model: $X = AS + N$, with $S \sim \mathcal{N}(0, P)$ and $N \sim \mathcal{N}(0, \Sigma)$. We find on the one hand

$$-\log(p(X|S,\theta)) = -\log(p(X - AS|S,\theta)) \tag{C.4}$$

$$= \log|\Sigma| + \mathrm{Tr}\left((R^{XX} - 2AR^{SX} + AR^{SS}A^\top)\Sigma^{-1}\right) \tag{C.5}$$

and on the other hand

$$-\log(p(S|\theta)) = \log|P| + \mathrm{Tr}\left(R^{SS}P^{-1}\right) \tag{C.6}$$

which gives

$$-\log(p(X,S|\theta)) = \mathrm{Tr}\left((R^{XX} - 2AR^{SX} + AR^{SS}A^\top)\Sigma^{-1}\right) + \mathrm{Tr}\left(R^{SS}P^{-1}\right) + \log|P| + \log|\Sigma|.$$

Then, in the Whittle approximation, the EM functional for a spectral model is simply the weighted sum of the previous EM functionals in each band, which gives the EM functional for SMICA:

$$\Phi = \sum_{b=1}^{B} n_b \left[\mathrm{Tr}((R_b^{XX} - 2AR_b^{SX} + AR_b^{SS}A^\top)\Sigma_b^{-1}) + \mathrm{Tr}(R_b^{SS}P_b^{-1}) + \log|\Sigma_b| + \log|P_b| \right] \quad , \tag{C.7}$$

which should be minimized with respect to the parameters θ.

- Optimizing P_b: the source powers are decoupled from the other parameters in (C.7). Minimization of Φ w.r.t. P_b is easily obtained by canceling the gradient, yielding:

$$P_b = \mathrm{diag}(R_b^{SS}) \ .$$

- Optimizing Σ_b: the mixing matrix A and the noise covariance are entangled in eq. (C.7), rendering the analytic minimization of Φ impossible. Therefore, we first minimize Φ w.r.t Σ_b, keeping A constant. This yields:

$$\Sigma_b = \mathrm{diag}(R_b^{XX} - 2AR_b^{SX} + AR_b^{SS}A^\top) \ .$$

- Optimizing A: keeping the noise levels fixed, minimizing Φ w.r.t. A yields, by canceling the gradient: $\sum_{i=1}^{B} n_b \Sigma_b^{-1}(R_b^{XS} - AR_b^{SS}) = 0$. This can be seen as a system of equations for the rows of A which, thanks to the diagonality of Σ_b, is easily seen to decouple across the rows. For each row, simple algebra yields the close form solution:

$$\text{for } \ r = 1 \dots p, \ \ A_{r:} = Q_{r:} M_r^{-1} \quad \text{with} \quad Q = \sum_{b=1}^{B} n_b \Sigma_b^{-1} R_b^{XS} \quad M_r = \sum_{b=1}^{B} n_b \sigma_{i,r}^{-2} R_b^{SS}.$$

  Therefore, the EM update of A only requires solving p linear systems of size $q \times q$.

Implementation details. The EM algorithm iterates the E and M step until a certain convergence criterion is reached. In practice, iterations are stopped when the difference between two consecutive values of the log-likelihood is below a threshold: $\mathcal{L}^{t+1} > \mathcal{L}^t - \varepsilon$. In order to have a good initialization for the

algorithm, we first fit the model with a fixed noise level for each bin: we estimate Σ subject to $\Sigma_b = \Sigma$ for all i. In this setting, the M-step is much simpler and computationally quicker. Then, the core SMICA algorithm with free noise starts with Σ_b all equal to the estimated noise level, and A and P_b start from the same initial value.

References

[1] Ablin, P., Cardoso, J.-F., Gramfort, A., 2018. Faster ICA under orthogonal constraint. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4464–4468.

[2] Ablin, P., Cardoso, J.-F., Gramfort, A., 2018. Faster independent component analysis by preconditioning with Hessian approximations. IEEE Transactions on Signal Processing 66 (15), 4040–4049.

[3] Ablin, P., Cardoso, J.-F., Gramfort, A., April 2019. Beyond Pham's algorithm for joint diagonalization. In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN).
URL https://hal.archives-ouvertes.fr/hal-01936887

[4] Anemller, J., Sejnowski, T. J., Makeig, S., 2003. Complex independent component analysis of frequency-domain electroencephalographic data. Neural Networks 16 (9), 1311 – 1323, neuroinformatics.
URL http://www.sciencedirect.com/science/article/pii/S0893608003002442

[5] Artoni, F., Delorme, A., Makeig, S., 2018. Applying dimension reduction to EEG data by principal component analysis reduces the quality of its subsequent independent component decomposition. NeuroImage 175, 176–187.

[6] Barthélemy, Q., Gouy-Pailler, C., Isaac, Y., Souloumiac, A., Larue, A., Mars, J. I., 2013. Multivariate temporal dictionary learning for EEG. Journal of neuroscience methods 215 (1), 19–28.

[7] Bell, A. J., Sejnowski, T. J., 1995. An information-maximization approach to blind separation and blind deconvolution. Neural computation 7 (6), 1129–1159.

[8] Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., Moulines, E., 1997. A blind source separation technique using second-order statistics. IEEE Transactions on signal processing 45 (2), 434–444.

[9] Brillinger, D. R., 2001. Time series: data analysis and theory. SIAM.

[10] Brookes, M. J., Woolrich, M., Luckhoo, H., Price, D., Hale, J. R., Stephenson, M. C., Barnes, G. R., Smith, S. M., Morris, P. G., 2011. Investigating the electrophysiological basis of resting state networks using magnetoencephalography. Proceedings of the National Academy of Sciences 108 (40), 16783–16788.

[11] Brown, R. G., Hwang, P. Y., et al., 1992. Introduction to random signals and applied Kalman filtering. Vol. 3. Wiley New York.

[12] Buzsáki, G., Draguhn, A., 2004. Neuronal oscillations in cortical networks. Science 304 (5679), 1926–1929.
URL https://science.sciencemag.org/content/304/5679/1926

[13] Cardoso, J.-F., Snoussi, H., Delabrouille, J., 2002. Blind separation of noisy Gaussian stationary sources. application to cosmic microwave background imaging. In: 2002 11th European Signal Processing Conference. IEEE, pp. 1–4.

[14] Cardoso, J.-F., Souloumiac, A., 1993. Blind beamforming for non-gaussian signals. IEE proceedings F (radar and signal processing) 140 (6), 362–370.

[15] Comon, P., 1994. Independent component analysis, a new concept? Signal processing 36 (3), 287–314.

[16] Congedo, M., Gouy-Pailler, C., Jutten, C., 2008. On the blind source separation of human electroencephalogram by approximate joint diagonalization of second order statistics. Clinical Neurophysiology 119 (12), 2677–2686.

[17] Dammers, J., Schiek, M., Boers, F., Silex, C., Zvyagintsev, M., Pietrzyk, U., Mathiak, K., Oct 2008. Integration of amplitude and phase statistics for complete artifact removal in independent components of neuromagnetic recordings. IEEE Transactions on Biomedical Engineering 55 (10), 2353–2362.

[18] Dehghani, N., Bédard, C., Cash, S. S., Halgren, E., Destexhe, A., Dec 2010. Comparative power spectral analysis of simultaneous elecroencephalographic and magnetoencephalographic recordings in humans suggests non-resistive extracellular media. Journal of Computational Neuroscience 29 (3), 405–421. URL https://doi.org/10.1007/s10827-010-0263-2

[19] Delabrouille, J., Cardoso, J.-F., Patanchon, G., 2003. Multidetector multicomponent spectral matching and applications for cosmic microwave background data analysis. Monthly Notices of the Royal Astronomical Society 346 (4), 1089–1102.

[20] Delorme, A., Palmer, J., Onton, J., Oostenveld, R., Makeig, S., 2012. Independent EEG sources are dipolar. PloS one 7 (2), e30135.

[21] Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 39 (1), 1–22.

[22] Doron, E., Yeredor, A., 2004. Asymptotically optimal blind separation of parametric Gaussian sources. In: International Conference on Independent Component Analysis and Signal Separation. Springer, pp. 390–397.

[23] Dyrholm, M., Makeig, S., Hansen, L. K., 2007. Model selection for convolutive ICA with an application to spatiotemporal analysis of EEG. Neural Computation 19 (4), 934–955.

[24] Gómez-Herrero, G., Atienza, M., Egiazarian, K., Cantero, J. L., 2008. Measuring directional coupling between EEG sources. Neuroimage 43 (3), 497–508.

[25] Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., Hmlinen, M., 2013. MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience 7, 267.
URL https://www.frontiersin.org/article/10.3389/fnins.2013.00267

[26] Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., Hämäläinen, M. S., 2014. MNE software for processing MEG and EEG data. Neuroimage 86, 446–460.

[27] Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., Lounasmaa, O. V., 1993. Magnetoencephalographytheory, instrumentation, and applications to noninvasive studies of the working human brain. Reviews of modern Physics 65 (2), 413.

[28] Hari, R., Salmelin, R., 1997. Human cortical oscillations: a neuromagnetic view through the skull. Trends in Neurosciences 20 (1), 44 – 49.
URL http://www.sciencedirect.com/science/article/pii/S0166223696100655

[29] Hunter, J. D., 2007. Matplotlib: A 2d graphics environment. Computing in science & engineering 9 (3), 90–95.

[30] Hyvärinen, A., 1998. Independent component analysis in the presence of Gaussian noise by maximizing joint likelihood. Neurocomputing 22 (1-3), 49–67.

[31] Hyvärinen, A., 1999. Fast and robust fixed-point algorithms for independent component analysis. IEEE transactions on Neural Networks 10 (3), 626–634.

[32] Hyvärinen, A., Karhunen, J., Oja, E., 2004. Independent component analysis. Vol. 46. John Wiley & Sons.

[33] Hyvärinen, A., Ramkumar, P., Parkkonen, L., Hari, R., 2010. Independent component analysis of short-time Fourier transforms for spontaneous EEG/MEG analysis. NeuroImage 49 (1), 257–271.

[34] Ikeda, S., Toyama, K., 2000. Independent component analysis for noisy dataMEG data analysis. Neural Networks 13 (10), 1063–1074.

[35] Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., Mckeown, M. J., Iragui, V., Sejnowski, T. J., 2000. Removing electroencephalographic artifacts by blind source separation. Psychophysiology 37 (2), 163–178.

[36] Lee, T.-W., Girolami, M., Sejnowski, T. J., 1999. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. Neural computation 11 (2), 417–441.

[37] Liu, F., Wang, S., Rosenberger, J., Su, J., Liu, H., 2017. A sparse dictionary learning framework to discover discriminative source activations in eeg brain mapping. In: Thirty-First AAAI Conference on Artificial Intelligence.

[38] Makeig, S., Bell, A. J., Jung, T.-P., Sejnowski, T. J., 1996. Independent component analysis of electroencephalographic data. In: Advances in neural information processing systems. pp. 145–151.

[39] Makeig, S., Debener, S., Onton, J., Delorme, A., 2004. Mining event-related brain dynamics. Trends in cognitive sciences 8 (5), 204–210.

[40] Mantini, D., Franciotti, R., Romani, G. L., Pizzella, V., 2008. Improving MEG source localizations: an automated method for complete artifact removal based on independent component analysis. NeuroImage 40 (1), 160–173.

[41] Niedermeyer, E., Lopes da Silva, F. H., 2005. Electroencephalography : basic principles, clinical applications, and related fields, 5th Edition. Philadelphia ; London : Lippincott Williams & Wilkins.

[42] Nikulin, V. V., Nolte, G., Curio, G., 2011. A novel method for reliable and fast extraction of neuronal EEG/MEG oscillations on the basis of spatio-spectral decomposition. NeuroImage 55 (4), 1528–1535.

[43] Olshausen, B. A., Field, D. J., 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381 (6583), 607–609.

[44] Palmer, J. A., Kreutz-Delgado, K., Makeig, S., 2012. AMICA: An adaptive mixture of independent component analyzers with shared components. Swartz Center for Computatonal Neursoscience, University of California San Diego, Tech. Rep.

[45] Parra, L., Spence, C., 2000. Convolutive blind separation of non-stationary sources. IEEE transactions on Speech and Audio Processing 8 (3), 320–327.

[46] Pham, D.-T., Cardoso, J.-F., Jan. 2003. Source adaptive blind source separation: Gaussian models and sparsity. In: Wavelets: Applications in Signal and Image Processing, X, Proc. of SPIE. Vol. 5207. San Diego, pp. 340 – 351.

[47] Pham, D. T., Garat, P., 1997. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. IEEE transactions on Signal Processing 45 (7), 1712–1725.

[48] Rocca, D. L., Zilber, N., Abry, P., van Wassenhove, V., Ciuciu, P., 2018. Self-similarity and multifractality in human brain activity: A wavelet-based analysis of scale-free brain dynamics. Journal of Neuroscience Methods 309, 175 – 187.
URL http://www.sciencedirect.com/science/article/pii/S0165027018302784

[49] Scherg, M., Von Cramon, D., Jan. 1985. Two bilateral sources of the late AEP as identified by a spatio-temporal dipole model. Electroencephalogr. Clin. Neurophysiol. 62 (1), 32–44.

[50] Stephen, J. M., Coffman, B. A., Jung, R. E., Bustillo, J. R., Aine, C., Calhoun, V. D., 2013. Using joint ICA to link function and structure using MEG and DTI in schizophrenia. Neuroimage 83, 418–430.

[51] Subasi, A., Gursoy, M. I., 2010. EEG signal classification using PCA, ICA, LDA and support vector machines. Expert systems with applications 37 (12), 8659–8666.

[52] Tang, A. C., Sutherland, M. T., McKinney, C. J., 2005. Validation of SOBI components from high-density EEG. NeuroImage 25 (2), 539–553.

[53] Urigüen, J. A., Garcia-Zapirain, B., 2015. EEG artifact removalstate-of-the-art and guidelines. Journal of neural engineering 12 (3), 031001.

[54] Vigário, R., Sarela, J., Jousmiki, V., Hamalainen, M., Oja, E., 2000. Independent component approach to the analysis of EEG and MEG recordings. IEEE transactions on biomedical engineering 47 (5), 589–593.

[55] Ziehe, A., Laskov, P., Nolte, G., Müller, K.-R., 2004. A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. Journal of Machine Learning Research 5 (Jul), 777–800.