NOTICE: this is the author’s version of a work that was accepted for publication in Signal Processing.
Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in Signal Processing, vol. 91, no. 8, Aug. 2011, pp. 1963-1972

Spatially constrained ICA algorithm with an application in EEG processing ☆

Maarten De Vos[a,b], Lieven De Lathauwer[a,c], Sabine Van Huffel[a]

[a] *Department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven, Leuven, Belgium*

[b] *Neuropsychology lab, Department of Psychology, University of Oldenburg, Oldenburg, Germany*

[c] *Katholieke Universiteit Leuven Campus Kortrijk, Group Science, Engineering and Technology, Kortrijk, Belgium*

Abstract

Independent Component Analysis (ICA) aims at blindly decomposing a linear mixture of independent sources. It has lots of applications in diverse research areas. In some applications, there is prior knowledge on the sources and/or the mixing vectors. This prior knowledge can be incorporated in the computation of the independent sources. In this paper we provide an algorithm for so-called spatially constrained ICA (scICA). The algorithm deals with the situation when one mixing vector is exactly known. Also the generalization to more mixing vectors is discussed. Numerical experiments

☆This research is funded by a PhD grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). Research Council KUL: GOA Ambiorics, GOA MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC), PFV/10/002 (OPTEC), IDO 05/010 EEG-fMRI, IDO 08/013 Autism, IOF-KP06/11 FunCopt; Flemish Government: FWO G.0302.07 (SVM), G.0341.07 (Data fusion), G.0427.10N (Integrated EEG-fMRI) research communities (ICCoS, ANMMM), FWO-G.0321.06 (Tensors/Spectral Analysis); TBM070713-Accelero, TBM070706-IOTA3, TBM080658-MRI (EEG-fMRI);IBBT

Email address: `Kasteelpark Arenberg 10, 3001 Heverlee, Belgium, maarten.devos@esat.kuleuven.be` (Maarten De Vos)

are reported that allow us to assess the improvement in accuracy that can be achieved with these algorithms compared to fully blind ICA and to a previously proposed constrained algorithm. We illustrate the approach with a biomedical application.
Keywords: Independent Component Analysis, prior knowledge, constrained ICA

1. Introduction

Independent Component Analysis (ICA) or Blind Source Separation (BSS) [8, 19, 12] is a technique that extracts statistically independent components from a set of measured signals. It has become a standard technique in neuroscience in order to decompose electro-encephalogram (EEG) and functional Magnetic Resonance Imaging (fMRI) recordings into individual brain activation components, see e.g. [31, 2]. Standard ICA has two limitations, which form the starting point for this paper. First, the extracted components are not ordered. When someone is only looking for the presence of one component, all the extracted components have to be visually inspected. Usually, complex features have to be computed in order to recognize the signal of interest. Second, in some applications, there is prior knowledge about the source signals being sought. ICA methods are blind-box techniques that do not exploit this available prior knowledge.

In order to improve the performance, several attempts have been made to incorporate the available signal characteristics and develop "semi-blind" source separation techniques. If one has information on the statistical dis-

tributions, one could incorporate prior knowledge in a Bayesian framework, see e.g. [30, 15, 24]. This has the disadvantage of having to specify all priors for the distribution of the model parameters. ~~Furthermore, no closed-form equations can be derived.~~ For the case that the source distributions are not fully known, but that some components have large second- or higher-order statistics, a so-called "robust BSS approach" has been introduced in [28]. (This should not be confused with "robust ICA" introduced in [34].) When only kurtosis signs are known, a BSS algorithm, proposed in [35], can be used. When information about the shape of one or more source signals is assumed, this prior knowledge can be used as a constraint in the ICA optimisation algorithm, leading to the introduction of temporally constrained ICA [13]. Following this idea, in e.g. [23, 14, 3] other optimisation algorithms are proposed to extract sources resembling a reference signal. Spectral information is incorporated in [27].

Besides trying to use information about the source signals, a second way to constrain ICA is to impose constraints on the columns of the mixing matrix. This is useful when the spatial distribution of one (or more) source(s) is (approximately) known. This is called spatially constrained ICA (scICA). In [11], a first algorithm was proposed. The article also contains a thorough discussion on the difference between hard and soft constraints and on the trade-off between imposing prior knowledge constraints and statistical independence constraints. ~~Indeed, in some situations inaccurate prior knowledge can conflict with assumptions of statistical independence.~~ Indeed, inaccurate spatial constraints may conflict with the assumption of statistical independence, especially if the latter is only approximately satisfied. See §1.1.5 for a

discussion of this point. In [21], it is mentioned (without giving details) that spatial constraints can be incorporated with Lagrange multipliers to solve an unconstrained problem.

This type of ICA with prior knowledge is potentially very important for applications. For instance, in wireless communications, the known mixing vector could correspond to a transmitter of which the position is known not to change from one transmitted block to the other. One can imagine situations in wireless communication where one or several users do not move. This could in principle be detected and periodically checked. In §4, we present an application in artifact removal. At the moment of writing, we discovered an application of semi-blind ICA in fMRI [22].

If a mixing vector is known, it may be tempting to work by deflation. That is, one could project the data on the known vector and continue with the residu. This approach is incorrect. It implies that the remaining mixing vectors lie in the orthogonal complement of the known mixing vector, which is not necessarily true. If the mixing matrix is orthogonal, i.e., after a pre-whitening, then the greedy approach may work. However, pre-whitening in the case where a mixing vector is known, is still an open problem, a fortiori when the problem is strictly overdetermined. In this paper we explain how this problem can be solved. We propose a new algorithm to incorporate prior knowledge into the popular ICA algorithms JADE [5] and SOBI [1]. The algorithm assumes that there is exact information about one or more mixing vectors.

The novelty of this work is threefold. First, we show how the prior knowledge should be taken into account in the pre-whitening step. Next, we develop a new algorithm that incorporates the prior knowledge in the separation of the pre-whitened data. We derive detailed closed-form formulas based on numerical algebra, and show that the algorithm outperforms the algorithm proposed in [11] if the spatial constraints are accurate. Third, we carefully evaluate in a simulation study the improvement in performance that can be obtained by incorporating prior knowledge compared to the fully blind approach.

This paper is organized as follows. First (§1.1.1), the necessary background on ICA algorithms, in particular the JADE and SOBI algorithms, is recalled. Second (§2), we develop our ideas about scICA. Two variants are discussed. A first algorithm computes the ICA decomposition when one mixing vector is exactly known (§2.2). Then we explain how to generalize the algorithm for the case where more mixing vectors are known (§2.3). The next section validates the methods in a simulation study (§3). We illustrate the technique by means of a real-life application, namely accurate eye artifact removal from EEG data (§4).

Our method is valid for both real- and complex-valued data. The presentation is in terms of complex-valued data.

1.1. Short Introduction to Algebraic Methods for ICA

1.1.1. General

Assume the basic linear statistical model

$$\mathbf{Y} = \mathbf{M} \cdot \mathbf{X} \quad (+\mathbf{N}) \tag{1}$$

where $\mathbf{Y} \in \mathbb{C}^{I \times J}$ is called the observation matrix, $\mathbf{X} \in \mathbb{C}^{P \times J}$ the source matrix and $\mathbf{N} \in \mathbb{C}^{I \times J}$ the additive Gaussian noise. $\mathbf{M} \in \mathbb{C}^{I \times P}$ is the mixing matrix. I is the number of sensors, P is the number of sources, and J the number of samples in each observation.

The goal of ICA is to estimate the mixing matrix $\mathbf{M}$ and/or the source matrix $\mathbf{X}$, given only the observation matrix $\mathbf{Y}$. In this work, we restrict our discussion to the so-called overdetermined case, i.e., $P \leqslant I$. For strategies to solve the underdetermined ICA problem, when there are more sources than sensors, we refer to [17, 16] and references therein.

Blind identification of $\mathbf{M}$ in (1) is only possible when particular assumptions about the sources are made. The most common assumption is that the sources are mutually statistically independent (implying diagonal covariance and higher-order cumulants), as well as independent from the noise components [7]. A computationally efficient algorithm, called Joint Approximate Diagonalization of Eigenmatrices (JADE) is described in [5]. An overview of other algorithms is given in [12, 8]. An other assumption that can be made, is that the sources are mutually uncorrelated but individually autocorrelated. In this case, an algorithm based on second-order statistics can be used to estimate the underlying sources, called Second Order Blind Identification (SOBI) [1].

In the following, we start with a note on pre-whitening, because we propose an alternative in §2.2.1 and §2.2.2. We also briefly review the idea of JADE and SOBI as starting point for §2.2.3. We conclude this introduction with a brief discussion of the possible conflict between spatial prior knowledge and the assumption of statistical independence.

1.1.2. Pre-whitening

From an algebraic point of view, saying that source signals $\mathbf{X}$ are uncorrelated means that the covariance matrix of $\mathbf{X}$, $\mathbf{C_X}$, is a diagonal matrix. Saying that the signals in $\mathbf{X}$ are independent also implies that all the higher-order cumulants of $\mathbf{X}$, $\mathcal{C}_{\mathbf{X}}^{(N)}$, are diagonal. The definition of higher-order cumulants and their properties can be found in [19, 25, 26].

With the linear model from equation (1), we can now write:

$$\mathbf{C_Y} = \mathbf{M} \cdot \mathbf{C_X} \cdot \mathbf{M}^H + \mathbf{C_N} \tag{2}$$

$$(\mathbf{C_Y})_{ij} = \sum_{k=1}^{P} m_{ik} (\mathbf{C_X})_{kk} m_{kj}^* \tag{3}$$

$$\begin{aligned} (\mathcal{C}_{\mathbf{Y}}^{(4)})_{ijkl} &= \sum_{m=1}^{P}\sum_{n=1}^{P}\sum_{o=1}^{P}\sum_{p=1}^{P} (\mathcal{C}_{\mathbf{X}}^{(4)})_{mnop} m_{im} m_{jn}^* m_{ko} m_{ln}^* \\ &+ (\mathcal{C}_{\mathbf{N}}^{(4)})_{ijkl} \end{aligned} \tag{4}$$

in which $\mathbf{C_X}$ and $\mathcal{C}_{\mathbf{X}}^{(4)}$ are diagonal and $\mathcal{C}_{\mathbf{N}}^{(4)}$ drops if the noise is Gaussian. Different kinds of algebraic ICA algorithms can be distinguished, depending on how equations (2) and (4) are combined. One approach is to solve the ICA problem using only higher-order statistics. For algebraic techniques, we refer to [4, 17, 18]. Equations (2) and (4) can also be combined in a more

or less balanced way [33, 20]. However, most algorithms start with a so-called pre-whitening step. In this pre-whitening step, as much information as possible is extracted from (2). In this stage, the mixing matrix can be identified up to a rotation factor. Afterwards, (4) has to be solved in some approximative manner. In what follows, we drop the noise terms for clarity.

In the pre-whitening step, the observation matrix $\mathbf{Y}$ is transformed into an other matrix of which the corresponding stochastic variables have unit covariance. One starts with the computation of the covariance matrix $\mathbf{C_Y}$ of the signal. We assume for convenience of notation that the signals are zero-mean. We have:

$$\mathbf{C_Y} = \frac{1}{J}\mathbf{Y} \cdot \mathbf{Y}^H \tag{5}$$

$$= \frac{1}{J}(\mathbf{M} \cdot \mathbf{X})(\mathbf{M} \cdot \mathbf{X})^H \tag{6}$$

$$= \mathbf{M} \cdot \mathbf{C_X} \cdot \mathbf{M}^H \tag{7}$$

$$= (\mathbf{US}) \cdot (\mathbf{US})^H \tag{8}$$

In the last step, we assume without loss of generality that the source signals have unit variance, and we replace $\mathbf{M}$ by its singular value decomposition $\mathbf{USV}^H$. A possible whitening matrix is then given by the pseudo-inverse of $\mathbf{US}$. We denote the pre-whitened observation matrix as

$$\mathbf{Y}_w = (\mathbf{US})^\dagger \cdot \mathbf{Y} \in \mathbb{C}^{P \times J}. \tag{9}$$

The pre-whitened mixing matrix is denoted by

$$\mathbf{M}_w = (\mathbf{US})^\dagger \cdot \mathbf{M} = \mathbf{V}^H. \tag{10}$$

In many applications of ICA (e.g., EEG, fMRI, ...), the number of sensors is higher than the number of sources. In the pre-whitening step, the dimensionality of the ICA problem is reduced by projecting $\mathbf{Y}$ onto the signal subspace. From the equations, it can be seen that the mixing matrix $\mathbf{M} = \mathbf{USV}^H$ can only be estimated from (8) up to a unitary matrix $\mathbf{V}$.

1.1.3. JADE

The assumption of mutually statistically independent sources means algebraically that all the higher-order cumulants of $\mathbf{X}$ are diagonal tensors. Equation (4) implies that the mixing matrix can be estimated as the matrix that diagonalizes the higher-order cumulant. This mixing matrix will be a unitary matrix when the observation matrix is first pre-whitened ($\mathbf{M}_w = \mathbf{V}^H$). An efficient algorithm to approximately diagonalize the cumulant is described in [5]. The cumulant slices of $\mathcal{C}^{(4)}_{\mathbf{Y}_w}$ can be rearranged into a matrix $\mathbf{C}^{(4)}_{\mathbf{Y}_w} \in \mathbb{C}^{P^2 \times P^2}$. It was shown that the mixing matrix can be estimated as the unitary matrix that approximately *simultaneously* diagonalizes the matrices that correspond to a basis of the column span of this matrix. The best choice for this set of matrices are the P dominant eigenmatrices of $\mathbf{C}^{(4)}_{\mathbf{Y}_w}$, which we will denote by $\mathbf{A}_p \in \mathbb{C}^{P \times P}, p = 1, \ldots, P$. So in order to estimate the mixing matrix, the P dominant eigenmatrices $\mathbf{A}_p$ of $\mathbf{C}^{(4)}_{\mathbf{Y}_w}$ have to be jointly (approximately) diagonalized. For the orthogonal simultaneous matrix diagonalization, one can resort to the Jacobi technique developed in [6, 5].

1.1.4. SOBI

When the sources are individually correlated in time, but mutually uncorrelated, an ICA algorithm based on second-order statistics can be derived [1].

Mathematically, this means that not only the covariance matrix of the sources is diagonal, but only that the covariance matrices at different time lags τ are diagonal. Under these assumptions, the mixing matrix after pre-whitening $\mathbf{M}_w$ is the unitary matrix that jointly diagonalizes all the covariance matrices:

$$\mathbf{C}_{\mathbf{Y}_w}(\tau) = \frac{1}{J}\mathbf{Y}_w(t)\mathbf{Y}_w(t+\tau)^H \tag{11}$$

$$= \mathbf{M}_w \cdot \mathbf{C}_{\mathbf{X}}(\tau) \cdot \mathbf{M}_w^H \qquad \forall\tau \tag{12}$$

These matrices can also be referred to as $\mathbf{A}_p, p = 0, \ldots, T$ (cf. JADE).

1.1.5. Spatial prior knowledge versus statistical independence

Equations (2) and (4) imply that, in the noise-free case, the observed covariance matrix $\mathbf{C}_\mathbf{Y}$ and fourth-order cumulant tensor $\mathcal{C}_\mathbf{Y}^{(4)}$ can be exactly diagonalized by the mixing matrix $\mathbf{M}$. The diagonality follows from the statistical independence of the sources. In the overdetermined case, the decomposition of the fourth-order cumulant tensor (and, a fortiori, the joint decomposition of the covariance matrix and the fourth-order cumulant tensor) is essentially unique if (i) at most one of the sources is non-kurtic and (ii) there are no collinear mixing vectors [8, Chapter 9]. This means that the exact mixing matrix can in principle be computed by decomposing the tensor (and the matrix). This is the foundation of all cumulant-based approaches to ICA. In this paper we assume that we have prior knowledge of one or several

mixing vectors. If this spatial prior knowledge is reliable, then it has to be consistent with the outcome of the decomposition.

On the other hand, assume that we know the full mixing matrix, i.e., we have full spatial prior knowledge. If now the assumption of statistical independence is reliable, then the given matrix will indeed diagonalize the observed covariance matrix and cumulant tensor.

However, if the spatial prior knowledge is inaccurate and/or the assumption of statistical independence only approximately satisfied, then there may be a conflict between the two constraints. If the sources are not really independent, then the observed covariance matrix and cumulant tensor will not be exactly diagonalized by the true mixing matrix. On the other hand, if the spatial prior knowledge is inaccurate, then the decomposition of the cumulant tensor (and the covariance matrix) will yield a mixing matrix that does not satisfy the spatial constraints.

In this paper we assume that at least the spatial prior knowledge is reliable. This is relevant for applications, see Sections 3 and 4.

2. Imposing spatial constraints

2.1. The algorithm proposed in [11]

In [11], a first algorithm to impose spatial constraints was proposed. The algorithm started with a standard pre-whitening (see Equation (8)). The columns of the mixing matrix were divided in constrained ($\mathbf{M}_c$) and unconstrained ($\mathbf{M}_u$) columns. After prewhitening, we have:

$$\mathbf{A}_c = (\mathbf{US})^\dagger \cdot \mathbf{M}_c, \tag{13}$$

$$\mathbf{A}_u = (\mathbf{US})^\dagger \cdot \mathbf{M}_u. \tag{14}$$

The matrix $\mathbf{M}_u$ is initialized with unit norm random vectors. The matrix $\mathbf{A}_c$ is first orthogonalized by means of Gram-Schmidt, yielding a matrix $\hat{\mathbf{A}}_c$. Note that the known mixing vectors are projected on the column space of $\mathbf{U}$. This means that the spatial constraints are only partially exploited. Indeed, after estimating $\mathbf{A}_c$, only the projection of the known mixing vectors on the column space of $\mathbf{U}$ can be retrieved and not the known mixing vectors themselves. Then, at each iteration, the following steps are performed:

1. Perform one iteration of an ICA algorithm (e.g. FastICA [12]) to update the columns of $\mathbf{A}_u$.
2. Orthogonalise $[\mathbf{A}_u\, \hat{\mathbf{A}}_c]$ by means of Gram-Schmidt.
3. Impose spatial constraints on the columns of $\hat{\mathbf{A}}_c$.
4. Orthogonalise $[\hat{\mathbf{A}}_c\, \mathbf{A}_u]$ by means of Gram-Schmidt.

Some comments are in order. It can be seen that step 2 is in fact redundant. Step 3 is redundant too, since $\hat{\mathbf{A}}_c$ does not change. The unitary matrix that is produced by the Gram-Schmidt procedure is not the closest to the original matrix in least-squares sense. Moreover, the order of the mixing vectors affects the result, while in the original formulation of the ICA problem the sources can be arbitrarily permuted.

2.2. The mixing vector of one source is exactly known.

We assume that we know one mixing vector K exactly. We first explain how we make sure that the mixing vector is fully contained in the signal subspace. We recall that in the algorithm of [11], this is not guaranteed. This first step is necessary for reducing the dimensionality of the problem.

Then we conduct a full pre-whitening. In the next step, we estimate the full mixing matrix.

2.2.1. Dimensionality reduction

We have to determine the subspace of dimension P that includes K and explains as much of the observed data as possible. Formally, we have to find a matrix $\tilde{\mathbf{U}} \in \mathbb{C}^{I \times P}$ with mutually orthonormal columns that maximises $||\tilde{\mathbf{U}}^H \mathbf{Y}||$ under the constraint that $K \in span(\tilde{\mathbf{U}})$. (Note that the maximisation is consistent with what happens in a classical prewhitening.) This can be done as follows:

1. Normalise K to unit length. The result will be the first column of $\tilde{\mathbf{U}}$, denoted by $\tilde{U}_1$. In this way, $K \in span(\tilde{\mathbf{U}})$.
2. Project the observation matrix $\mathbf{Y}$ on the orthogonal complement of $\tilde{U}_1$ in order to obtain $\mathbf{Y}_2$. We have $\mathbf{Y}_2 = (\mathbf{I} - \tilde{U}_1 \tilde{U}_1^H) \cdot \mathbf{Y}$.
3. Define $\tilde{U}_2$ to $\tilde{U}_P$ as the $P-1$ dominant left singular vectors of $\mathbf{Y}_2$. These vectors allow us to maximise $||\tilde{\mathbf{U}}^H \mathbf{Y}||$.

In order to reduce the dimensionality, we now project the data matrix on the subspace that we have formed:

$$\mathbf{Y}_s = \tilde{\mathbf{U}}^H \cdot \mathbf{Y}, \tag{15}$$

where $\mathbf{Y}_s \in \mathbb{C}^{P \times J}$. (The subscript refers to "signal subspace".) The mixing matrix in the P-dimensional space is denoted by $\mathbf{M}_s$. We have

$$\mathbf{M} = \tilde{\mathbf{U}} \cdot \mathbf{M}_s. \tag{16}$$

After the dimensionality reduction step, we know the column space of the mixing matrix. This is the column space of $\tilde{\mathbf{U}}$. Note that the columns of $\tilde{\mathbf{U}}$

are not the dominant left singular vectors of $\mathbf{Y}$. Consequently, the projection (15) does not necessarily make the components of $\mathbf{Y}_s$ uncorrelated, as in unconstrained ICA.

2.2.2. Pre-whitening

The problem that remains to solve is a square $(P \times P)$ ICA problem. We start with a pre-whitening:

$$\mathbf{C}_{\mathbf{Y}_s} = \mathbf{Y}_s \cdot \mathbf{Y}_s^H \tag{17}$$

$$= \mathbf{M}_s \cdot \mathbf{C}_{\mathbf{X}} \cdot \mathbf{M}_s^H \tag{18}$$

$$= (\mathbf{QS}) \cdot (\mathbf{QS})^H, \tag{19}$$

in which the last equation shows the Eigenvalue Decomposition (EVD) of $\mathbf{C}_{\mathbf{Y}_s}$. The singular value decomposition of the mixing matrix after dimensionality reduction is given by:

$$\mathbf{M}_s = \mathbf{QSV}^H, \tag{20}$$

in which the unitary matrix $\mathbf{V}$ is still unknown. $\mathbf{Q}$ and $\mathbf{S}$ are obtained from the EVD (19). The overall pre-whitening matrix $\mathbf{M}_w \in \mathbb{C}^{I \times P}$ that takes into account the spatial prior knowledge, is given by

$$\mathbf{M}_w = \tilde{\mathbf{U}} \cdot \mathbf{QS}. \tag{21}$$

The pre-whitened data are denoted by

$$\mathbf{Y}_w = \mathbf{M}_w^\dagger \cdot \mathbf{Y} = \mathbf{S}^\dagger \mathbf{Q}^H \tilde{\mathbf{U}}^H \cdot \mathbf{Y} \in \mathbb{C}^{P \times J}. \tag{22}$$

The pre-whitened observations are related to the sources as follows:

$$\mathbf{Y}_w = \mathbf{V}^H \cdot \mathbf{X}. \tag{23}$$

2.2.3. Determination of $\boldsymbol{V}$ using prior knowledge

The unitary matrix $\mathbf{V} \in \mathbb{C}^{P \times P}$ can be determined as follows in order to estimate the full mixing matrix:

1. We know the mixing vector in the original I-dimensional space, K, and its normalised variant $\tilde{U}_1$. Consequently, we know the first column of $\mathbf{M}_s$ after dimensionality reduction, which we will denote as $M_{s,1}$. From equation (20) we then also know the first column of $\mathbf{V}^H$, denoted as V_1. The procedure is as follows. From equations (1) and (15), we have that $M_{s,1} = \tilde{\mathbf{U}}^H \tilde{U}_1 = E_1$, in which E_1 is the first canonical unit vector with 1 as its first entry and 0 as the other entries. ¿From equation (20), we have that

$$M_{s,1} = \mathbf{QS} \cdot V_1. \tag{24}$$

   Consequently,

$$V_1 = \mathbf{S}^\dagger \mathbf{Q}^H \cdot M_{s,1}. \tag{25}$$

   V_1 is subsequently normalized to unit length, which we can do because of the ICA scaling ambiguity.

2. The matrix $\mathbf{V}$ is unitary. By definition, its last $(P-1)$ columns $V_2, \ldots, V_P$ are orthogonal to V_1. Let $\mathbf{V}_1^\perp \in \mathbb{C}^{P \times (P-1)}$ denote an arbitrary column-wise orthonormal matrix of which the columns form an orthonormal basis for the orthogonal complement of V_1. Then we have $[V_2 \ldots V_P] = \mathbf{V}_1^\perp \cdot \mathbf{W}$, in which $\mathbf{W} \in \mathbb{C}^{(P-1)\times(P-1)}$ is a unitary matrix. Our goal is the estimation of $\mathbf{W}$. Let $\tilde{\mathbf{X}} \in \mathbb{C}^{(P-1)\times J}$ be the matrix obtained from $\mathbf{X}$ after dropping the first row, i.e., $\tilde{\mathbf{X}}$ contains all the sources except from the one of which the mixing vector is known.

Projection of $\mathbf{Y}_w$ on the orthogonal complement of V_1 now yields:

$$\tilde{\mathbf{Y}}_w = (\mathbf{V}_1^{\perp})^H \mathbf{Y}_w = \mathbf{W} \cdot \tilde{\mathbf{X}}. \tag{26}$$

Equation (26) is a classical ICA problem after pre-whitening, without constraints on the mixing matrix due to prior knowledge. This problem may be solved by means of, e.g., standard JADE or SOBI.

Summarizing our derivation, our estimate of the mixing matrix $\mathbf{M}$ is given by

$$\hat{\mathbf{M}} = \tilde{\mathbf{U}} \cdot \mathbf{QS} \cdot [V_1 \, \mathbf{V}_1^{\perp}] \cdot \begin{bmatrix} 1 & 0 \\ 0 & \mathbf{W} \end{bmatrix}. \tag{27}$$

2.3. The mixing vectors of several sources are exactly known.

We now assume that L vectors are known.

2.3.1. Dimensionality reduction

The first step in section 2.2.1 is replaced by the computation of an orthonormal basis of the span of the L known vectors. In the second step, we project on the orthogonal complement of this basis. In the third step, we extract $P - L$ singular vectors. This yields a column-wise orthonormal $\tilde{\mathbf{U}} \in \mathbb{C}^{I \times P}$ which we use for dimensionality reduction as in (15).

2.3.2. Pre-whitening

The pre-whitening can proceed as in section 2.2.2. Equation (20) remains valid.

2.3.3. Determination of $\boldsymbol{V}$ *using prior knowledge*

Like in equation (25), we have that

$$V_l = \mathbf{S}^{\dagger}\mathbf{Q}^H M_{s,l} \qquad l = 1, \ldots, L. \tag{28}$$

These vectors are not necessarily orthogonal. We will impose the statistical independence constraint while staying as closely as possible to the known vectors. Consequently, we replace the vectors $V_l, l = 1, \ldots, L$, by column-wise orthonormal vectors $Z_l, l = 1, \ldots, L$, that are as close as possible in the least squares sense, up to an irrelevant scaling. We stack the vectors V_l in $\mathbf{V}_{1:L} \in \mathbb{C}^{P \times L}$. Formally, we try to estimate the column-wise orthonormal matrix $\mathbf{Z} \in \mathbb{C}^{P \times L}$ and the diagonal matrix $\mathbf{D}$ that minimise

$$||\mathbf{V}_{1:L} \cdot \mathbf{D} - \mathbf{Z}||^2.$$

This can be done as follows [29]:

1. Normalise the columns of $\mathbf{V}_{1:L}$ to unit length.
2. Iterate until convergence:

   a) The computation of $\mathbf{Z}$, given $\mathbf{D}$, is an orthogonal Procrustes problem [9]. This means that, if the SVD of $\mathbf{V}_{1:L} \cdot \mathbf{D}$ is given by $\mathbf{ACB}^H$, then the optimal $\mathbf{Z}$ is given by $\mathbf{AB}^H$.

   b) The computation of $\mathbf{D}$, given $\mathbf{Z}$, is a standard linear least squares problem. We have: $\mathbf{D} = (diag(\mathbf{Z}^H \cdot \mathbf{V}_{1:L}))^{-1}$.

Note that this procedure is not an ordinary orthonormalisation of the vectors spanning a subspace, as proposed in [11], but an algorithm that guarantees that the vectors stay as closely as possible to the known vectors.

We can now proceed as in section 2.2.3 and reduce the problem to an unconstrained $((P-L)\times(P-L))$ ICA problem. The algorithm is summarised in Algorithm 1.

Algorithm 1 scICA, Case: mixture $\mathbf{Y}$ containing P sources of which L mixing vectors M_l are known.

1: Compute an orthonormal basis of the known mixing vectors and stack it in $\tilde{\mathbf{U}}_{1:L}$.
2: Project the data on the orthogonal complement of this basis. $\mathbf{Y}_2 = (\mathbf{I} - \tilde{\mathbf{U}}_{1:L}\tilde{\mathbf{U}}_{1:L}^H)\cdot\mathbf{Y}$.
3: Compute the $P-L$ dominant left singular vectors of $\mathbf{Y}_2$ and stack them in $\tilde{\mathbf{U}}_{(L+1):P}$.
4: Project the data on $\tilde{\mathbf{U}} = [\tilde{\mathbf{U}}_{1:L}\ \tilde{\mathbf{U}}_{(L+1):P}] : \mathbf{Y}_s = \tilde{\mathbf{U}}^H\cdot\mathbf{Y}$.
5: Whiten the data: $\mathbf{Y}_s = \mathbf{QSV}^H \Rightarrow \mathbf{Y}_w = \mathbf{S}^\dagger\mathbf{Q}^H\cdot\mathbf{Y}_s$.
6: Whiten the known vectors: $V_l = \mathbf{S}^\dagger\mathbf{Q}^H\tilde{\mathbf{U}}^H\cdot M_l$, $l = 1,\ldots,L$.
7: Normalise V_l to unit length, $l = 1,\ldots,L$.
8: Iterate until convergence:

a) $\mathbf{V}_{1:L}\cdot\mathbf{D} = \mathbf{ACB}^H \Rightarrow \mathbf{Z} = \mathbf{AB}^H$.

b) $\mathbf{D} = (diag(\mathbf{Z}^H\cdot\mathbf{V}_{1:L}))^{-1}$.

9: Compute the orthogonal complement $\mathbf{V}_{1:L}^\perp$ of $\mathbf{V}_{1:L}$ and project $\mathbf{Y}_w$ on it: $\tilde{\mathbf{Y}}_w = (\mathbf{V}_{1:L}^\perp)^H\mathbf{Y}_w$.
10: Solve a classical $((P-L)\times(P-L))$ ICA problem: $\tilde{\mathbf{Y}}_w = \mathbf{W}\cdot\tilde{\mathbf{X}}$.
11: The mixing matrix can be estimated as:
$\hat{\mathbf{M}} = \tilde{\mathbf{U}}\cdot\mathbf{QS}\cdot[V_{1:L}\ \mathbf{V}_{1:L}^\perp]\cdot\begin{bmatrix}\mathbf{I}_{L\times L} & 0\\ 0 & \mathbf{W}\end{bmatrix}$.
12: The sources can be estimated as $\hat{\mathbf{X}} = (\hat{\mathbf{M}})^\dagger\cdot\mathbf{Y}$.

3. Simulation study

In this section, we show the performance of the proposed algorithm for four synthetic datasets, inspired by applications in wireless communication. Namely, (1) is the standard base-band model for narrow-band sources. It is also the model that applies to line-of-sight propagation [32]. We compare our algorithm with unconstrained ICA and with the constrained algorithm, proposed in [11]. For a fair comparison, the higher-order steps of the different

algorithms are based on the JADE algorithm.

The independent sources are drawn from a binary distribution (1 or -1), with an equal probability of both values. These sources form the rows of $\mathbf{X}$. We consider 1000 samples of each source. Observation matrices $\tilde{\mathbf{Y}} \in \mathbb{C}^{6\times 1000}$ are generated in the following way:

$$\tilde{\mathbf{Y}} = \frac{\mathbf{Y}}{||\mathbf{Y}||_F} + \sigma_N \frac{\mathbf{N}}{||\mathbf{N}||_F}, \tag{29}$$

in which $\mathbf{Y} = \mathbf{M} \cdot \mathbf{X}$ exactly satisfies a linear mixture model with $P = 4$ independent sources and in which $\mathbf{N}$ represents zero-mean Gaussian noise.

We conducted Monte Carlo simulations consisting of 3500 runs. We evaluated the performance of the different algorithms by means of the Bit Error Rate (BER). The rows of the estimated sources $\hat{\mathbf{X}}$ were optimally ordered and scaled.

We considered five datasets. The entries of the mixing matrix $\mathbf{M}$ are drawn from a complex zero-mean unit-variance Gaussian distribution. White Gaussian noise was added. In the first dataset (Fig. 1), the variance of all sources was the same. We compared standard ("blind") JADE to scICA with one mixing vector perfectly known. We also plot the performance of the algorithm in [11], when one mixing vector was known. The figure shows that having perfect knowledge of one mixing vector mildly improves the estimation of the sources compared to standard JADE. There is no difference between the two constrained algorithms.

In the second simulation, we generated mixtures in the same way, but we gave the first source a ten times weaker contribution. This situation corresponds to a mixing matrix with a higher condition number. The prior knowledge concerned the weak source. From Fig. 2, it can be seen that

knowing the mixing vector improves the estimation of the sources a lot in this case. At high noise levels, some sources are not well estimated by standard ICA. By using prior knowledge, these sources can be estimated much more accurately. Knowing three mixing vectors further slightly improves the accuracy. The figures also illustrate that the new algorithm performs better than the algorithm in [11].

For the same dataset, we also estimated the BER when we have prior knowledge on an other mixing vector. The results are shown in Fig. 3. Even when we have prior knowledge about the mixing vector of a strong source, the BER decreases significantly.

In a third simulation (Fig. 4), we generated data $\mathbf{Y} \in \mathbb{C}^{8\times 1000}$ in the same way as the first simulation, but decreased the condition of the mixing matrix, by dividing the smallest singular value by a factor ten. Also in this simulation, exploiting prior knowledge about one or three mixing vectors decreases the BER.

In the fourth simulation (Fig. 5), we consider mixtures $\tilde{\mathbf{Y}} \in \mathbb{C}^{8\times 1000}$ containing five sources. Colored noise was generated by multiplying white noise with a matrix containing elements drawn from a Gaussian distribution. The color of the noise is unknown, and hence cannot be exploited. The figure shows that knowing one mixing vector improves the estimation compared to blind JADE. Knowing more mixing vectors further improves the estimation. The algorithm in [11] performs even worse than blind JADE. For the same dataset, we plot in Fig. 6 the canonical angle between the true and the estimated subspace. ¿From the figure it can be seen that the better performance of scICA is partly due to an improved estimation of the signal subspace in

the pre-whitening step. The algorithm in [11] estimates the signal subspace in the same way as blind JADE. In scICA the spatial prior knowledge is used to obtain a better estimate.

In the last simulation (Fig. 7), we estimate the error when prior knowledge is used that is not fully reliable. We generated mixtures $\tilde{\mathbf{Y}} \in \mathbb{C}^{8\times 1000}$ containing five sources. We decreased the condition number of the mixture in the same way as in the third simulation. The noise was white Gaussian. For the spatial constraints we used vectors $\tilde{M}_i$ generated as follows:

$$\tilde{M}_i = \frac{M_i}{||M_i||} + \alpha \frac{P_i}{||P_i||} \; i = 1, 2, \tag{30}$$

with P_i a vector in the orthogonal complement of M_i. So the mixing vectors used as prior information are only approximations of the true mixing vectors. The considered SNR was 25dB. The figure shows that scICA outperforms a fully blind decomposition up to deviations of 5 %. It turns out that scICA requires accurate spatial information. This is not unexpected, since imposing inaccurate prior knowledge as a hard constraint can be worse than not imposing prior knowledge at all.

These simulations show that the accuracy can be increased by incorporating spatial prior knowledge, at least when this prior knowledge is accurate. Furthermore, as we mentioned in the introduction, the components obtained from standard ICA are not ordered and all have to be inspected, even if just one is of interest. ICA with prior knowledge puts the signal(s) of interest in the first component(s). Postprocessing is not needed anymore, at least when the components of interest correspond to the components for which prior knowledge is available.

4. Application in EEG signal processing: eye blink removal

We now show an application of scICA in neuroscience, namely the removal of eye blink artifacts from EEG recordings. A full validation is outside the scope of this manuscript. We concatenated five EEG epochs of rest EEG contaminated with eye blinks. We ran standard ICA (SOBI) on this dataset. One source corresponds to the eye blink (see e.g. [10]). This source can be visually selected and the corresponding mixing vector retrieved. The mixing vector gives a robust estimate of the spatial distribution of the eye blink. There is very little variation between eye artifacts from the same person, because the eye balls do not move with respect to the electrodes. The known vector can then be used in SOBI-based scICA to remove *other* eye blinks in other parts of the data, where they would less easily have been found by blind SOBI. Indeed, although in many cases standard ICA gives a good separation between eye artifacts and brain activity, this is not always the case. When the spatial distribution of prominent brain activity, e.g. an epileptic seizure, overlaps with the eye blink and the time course of the blink is synchronised with the seizure activity, standard ICA can fail as we illustrate now.

Fig. 8 shows the original EEG dataset containing an epileptic seizure (mainly present on electrodes Fp1, F7 and F3) and eye blinks around time $= 0.5, 3.3 \,\text{and}\, 4.5s$. Fig. 9 shows the EEG when the artifact is removed by imposing in scICA that the first component is related to eye blinks, and afterwards subtracting this component. Fig. 10 shows the EEG cleaned by means of blind SOBI. In the latter figure, the blink is not fully removed at channels Fp1, F7 and F3. In this particular application, the method in [11] gave comparable results.

5. Discussion

Standard ICA blindly separates a linear mixture of statistically independent source signals. In several applications, one has prior knowledge about sources and/or mixing vectors. Hence, the development of semi-blind methods that take this prior knowledge into account is of interest. In this paper we propose a new semi-blind version of JADE and SOBI. The method works for both real- and complex-valued data, and an arbitrary number of spatial constraints.

Simulation results showed (Fig. 1) that in the case of well-conditioned data, scICA performs mildly better than blind ICA. When the mixing matrix is ill-conditioned (Fig. 2), scICA largely outperforms standard ICA. There is also a performance improvement in the case of noise with unknown color (Fig. 3). We also compared our algorithm with the algorithm proposed in [11]. The new algorithm never performs worse than the algorithm proposed in [11] and is clearly better in the presence of colored noise and when there is knowledge about more than one mixing vector. A limitation of the technique is that the prior knowledge has to be accurate. Further, scICA has the advantage over blind ICA that the sources for which spatial information is available, do not appear in a random order.

An interesting application is eye artifact removal from EEG recordings. A full validation study was outside the scope of this paper. However, the given example illustrates a limitation of standard ICA, that can be overcome by incorporating prior knowledge.

Our derivation is a necessary first step towards convolutive scICA, i.e., blind deconvolution with prior knowledge of one or more of the mixing filters. We believe that convolutive scICA is a promising conceptually new approach to blind multi-user access in wireless communication, when the position of one or more transmitters does not substantially change over time. The generalization is nontrivial and will be reported in a later paper.

References

[1] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second-order Statistics. IEEE Trans. Signal Processing, 45:434–444, 1997.

[2] V. Calhoun and T. Adali. Unmixing fMRI with independent component analysis. IEEE Engineering in Medicine and Biology Magazine, 25:79–90, 2006.

[3] V. D. Calhoun, T. Adali, M. Stevens, K. A. Kiehl, and J. J. Pekar. Semi-blind ICA of fMRI: A method for utilizing hypothesis-derived time courses in a spatial ICA analysis. NeuroImage, 35:527–538, 2005.

[4] J.-F. Cardoso. Iterative techniques for blind source separation using only fourth-order cumulants. In Proc. of the European Signal Processing Conference (EUSIPCO '92), pages 739–742, Brussels, Belgium, 1992.

[5] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non-Gaussian signals. IEE Proc. F, 140:362–370, 1994.

[6] J.-F. Cardoso and A. Souloumiac. Jacobi angles for simultaneous diagonalization. SIAM J. Matrix Anal. Appl., 17:161–164, 1996.

[7] P. Comon. Independent component analysis, a new concept? Signal Processing, 36:287–314, 1994.

[8] P. Comon and C. Jutten. Handbook of Blind Source Separation, Independent Component Analysis and Applications. Academic Press, 2010.

[9] G.H. Golub and C.F. Van Loan. Matrix computations. Third Edition The Johns Hopkins University Press, 1996.

[10] G. Gómez-Herrero, W. De Clercq, H. Anwar, O. Kara, K. Egiazarian, S. Van Huffel, and W. Van Paesschen. Automatic removal of ocular artifacts in the EEG without a reference EOG channel. In 7th Nordic Signal Processing Symposium (NORSIG), Reykjavik, Iceland, 2006.

[11] C.W. Hesse and C.J. James. On semi-blind source separation using spatial constraints with applications in EEG analysis. IEEE Trans. Biomed. Eng., 53:2525–2534, 2006.

[12] A. Hyvärinen, J. Karhunen and E. Oja. Independent Component Analysis. John Wiley & Sons, 2001.

[13] C.J. James and O. J. Gibson. Temporally constrained ICA: An application to artifact rejection in electromagnetic brain signal analysis. IEEE Trans. Biomed. Eng., 50:1108–1116, 2003.

[14] M. Jing and S. Sanei. A new constrained topographic independent component analysis for separation of epileptic seizure signals. Computational Intelligence and Neuroscience, id 21315, 2007.

[15] K.H. Knuth. Informed source separation: A Bayesian tutorial. In Proc. of the 13th European Signal Processing Conference (EUSIPCO '05), Antalya, Turkey, 2005.

[16] L. De Lathauwer and J. Castaing. Blind identification of underdetermined mixtures by simultaneous matrix diagonalization. IEEE Trans. Signal Processing, 56:1096–1105, 2008.

[17] L. De Lathauwer, J. Castaing, and J.-F. Cardoso. Fourth-order cumulant based underdetermined independent component analysis. IEEE Trans. Signal Processing, 55:2965–2973, 2007.

[18] L. De Lathauwer, B. De Moor, and J. Vandewalle. Independent component analysis based on higher-order-only ICA. In Proc. of 8th IEEE SP workshop on Statistical Signal and Array Processing, pages 356–359, Corfu, Greece, 1996.

[19] L. De Lathauwer, B. De Moor, and J. Vandewalle. An introduction to independent component analysis. Journal of Chemometrics, 14:123–149, 2000.

[20] L. De Lathauwer and J. Vandewalle. Dimensionality reduction in higher-order signal processing and rank-$(R_1, R_2, \ldots, R_N)$ reduction in multilinear algebra. Linear Algebra and its Applications, 391:31–55, 2004.

[21] M.A. Latif, S. Sanei, J. Chambers, and L. Shoker. Localization of abnormal EEG sources using blind source separation partially constrained by the locations of known sources. IEEE Signal Processing Letters, 13:117–120, 2006.

[22] Q.-H. Lin, J. Kiu, Y.-R. Zheng, H. Liang, and V. Calhoun. Semiblind spatial ICA of fMRI using spatial constraints. Human Brain Mapping, accepted, 2010.

[23] W. Lu and J.C. Rajapakse. Approach and applications of constrained ICA. IEEE Trans. on neural networks, 16:203–212, 2005.

[24] A. Mohamed-Djafari. Bayesian Source Separation: beyond PCA and ICA. In Proc. of European Symposium on Artificial Neural Networks (ESANN '06), Bruges, Belgium, 2006.

[25] C.L. Nikias and J. Mendel. Signal processing with higher-order spectra. IEEE Signal Processing Magazine, pages 10–37, 1993.

[26] C.L. Nikias and A.P. Petropulu. Higher-Order Spectra Analysis. A Nonlinear Signal Processing Framework. Englewood Cliffs, N.J.: Prentice-Hall, 1993.

[27] R. Phlypo, V. Zarzoso, and I. Lemahieu. Source extraction by maximising the variance in the conditional distribution tails. IEEE Trans. Signal Process., 58:305–316, 2010.

[28] M. Sahmoudi and K. Abed-Meraim. Robust blind separation algorithms for heavy-tailed sources. In Proc. of the IEEE Symposium on

Signal Processing and Information Technology (ISSPIT '04), pages 56–59, Rome, Italy, 2004.

[29] M. Signoretto, K. Pelckmans, L. De Lathauwer, and J.A.K. Suykens. Data-Dependent Norm Adaptation for Sparse Recovery in Kernel Ensembles Learning. Technical Report 09-97, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2009.

[30] H. Snoussi and A.M. Djafari. Bayesian unsupervised learning for source separation with mixture of gaussians prior. J. VLSI Signal Process. Syst., 37:263–279, 2004.

[31] A.C. Tang, M.T. Sutherland, and C.J. McKinney. Validation of SOBI components from high-density EEG. NeuroImage, 25:539–553, 2005.

[32] A.J. van der Veen. Algebraic methods for deterministic blind beamforming. Proc. of the IEEE, 86:1987 – 2008, 1998.

[33] A. Yeredor. Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. IEEE Trans. Signal Processing, 50:1545–1553, 2002.

[34] V. Zarzoso and P. Comon. Robust independent component analysis for blind source separation and extraction with application in electrocardiography. In 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE EMBS '08), pages 3344–3347, Vancouver, Canada, 2008.

[35] V. Zarzoso, R. Phlypo, and P. Comon. A Contrast for Independent

Component Analysis With Priors on the Source Kurtosis Signs. IEEE Signal Processing Letters, 15:501–504, 2008.

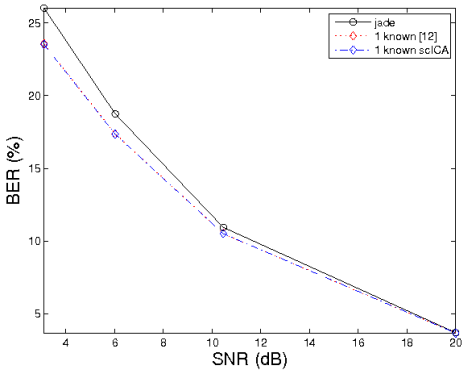

Figure 1: Results for the first dataset. The mean value of BER as a function of the SNR (dB) for unconstrained JADE (solid), for the scICA algorithm with one mixing vector known (dash-dotted), and for the algorithm in [11] with one mixing vector known (dotted). All sources have the same variance. The curves for the constrained algorithms coincide.

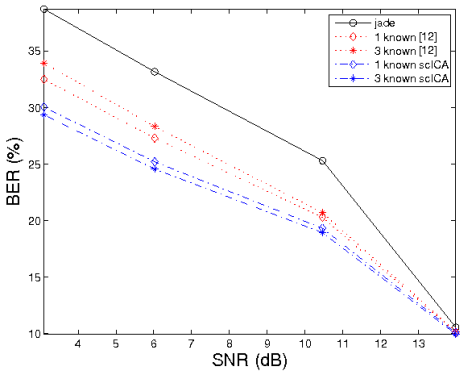

Figure 2: Results for the second dataset. The mean value of BER as a function of the SNR (dB) for unconstrained JADE (solid), for the scICA algorithm with one (triangle) and three (star) mixing vectors known (dash-dotted), and for the algorithm in [11] with one (triangle) and three (star) mixing vectors known (dotted). One source has a ten times smaller variance than the other sources. The prior knowledge concerned the mixing vector of this source.

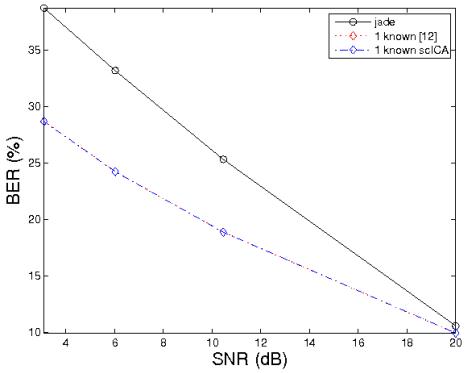

Figure 3: Results for the second dataset (continued). The mean value of BER as a function of the SNR (dB) for unconstrained JADE (solid), for the scICA algorithm with one mixing vector known (dash-dotted), and for the algorithm in [11] with one mixing vector known (dotted). The curves from the constrained algorithms coincide. One source has a ten times smaller variance than the other sources. The prior knowledge did not concern the mixing vector of this source.

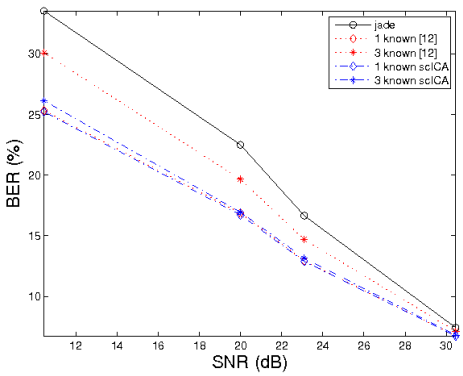

Figure 4: Results for the third dataset. The mean value of BER as a function of the SNR (dB) for unconstrained JADE (solid), for the scICA algorithm with one (triangle) and three (star) mixing vectors known (dash-dotted), and for the algorithm in [11] with one (triangle) and three (star) mixing vectors known (dotted).

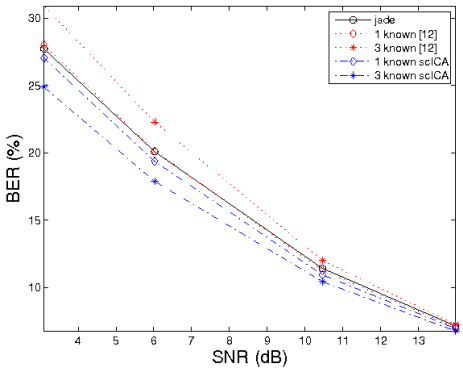

Figure 5: Results for the fourth dataset. The mean value of BER as a function of the SNR (dB) for unconstrained JADE (solid), for the scICA algorithm with one (triangle) and three (star) mixing vectors known (dash-dotted), and for the algorithm in [11] with one (triangle) and three (star) mixing vectors known (dotted). The noise was colored.

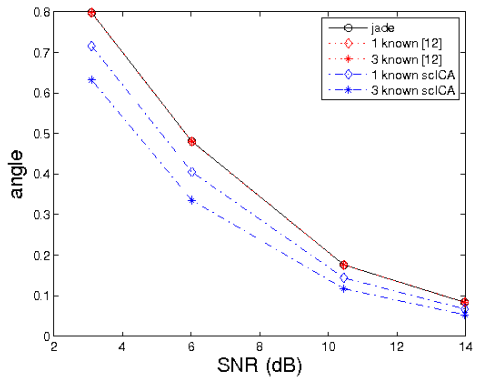

Figure 6: The canonical angle between the true subspace and the subspace estimated by unconstrained JADE (solid), the scICA algorithm with one (triangle) and three (star) mixing vectors known (dash-dotted), and the algorithm in [11] with one (triangle) and three (star) mixing vectors.

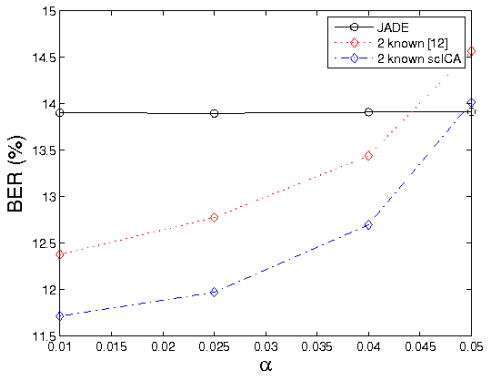

Figure 7: Results for the sixth dataset. The mean value of BER for SNR = 25 dB as a function of α, which reflects the inaccuracy of the prior knowledge for unconstrained JADE (solid), for the scICA algorithm with two mixing vectors approximately known (dash-dotted), and for the algorithm in [11] (dotted) with two mixing vectors known. The noise was white Gaussian.

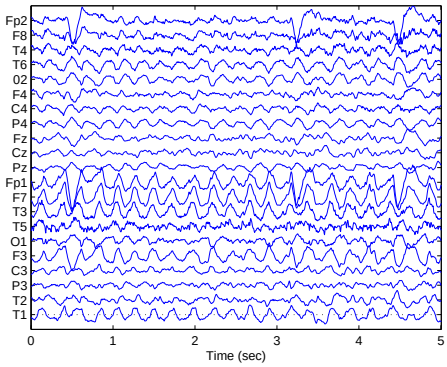

Figure 8: The original EEG containing an eye blink around time 0.5, 3.3 and 4.5 s.

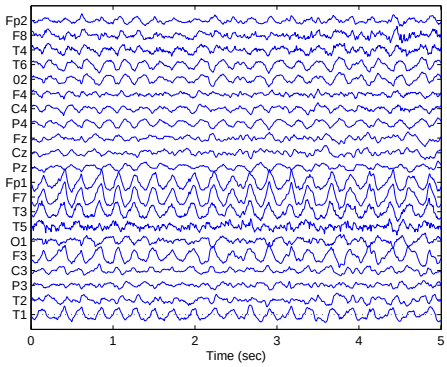

Figure 9: The EEG after eye artifact removal with scICA.

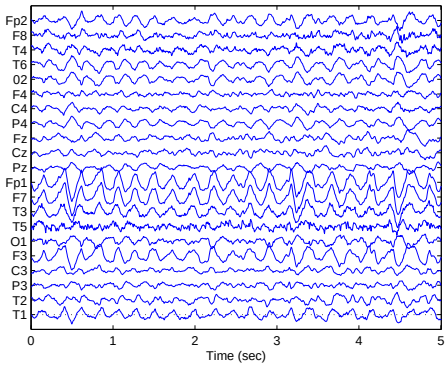

Figure 10: The EEG after eye artifact removal with blind SOBI.