Sparse Component Analysis: a New Tool for Data Mining

Pando Georgiev[1], Fabian Theis[2], Andrzej Cichocki[3], and Hovagim Bakardjian[3]

[1] ECECS Department, University of Cincinnati
Cincinnati, OH 45221 USA
pgeorgie@ececs.uc.edu
[2] Institute of Biophysics, University of Regensburg
D-93040 Regensburg, Germany
fabian@theis.name
[3] Brain Science Institute, RIKEN, Wako-shi, Japan
{cia,hova}@bsp.brain.riken.go.jp

Summary. In many practical problems for data mining the data $\mathbf{X}$ under consideration (given as $(m \times N)$-matrix) is of the form $\mathbf{X} = \mathbf{AS}$, where the matrices $\mathbf{A}$ and $\mathbf{S}$ with dimensions $m \times n$ and $n \times N$ respectively (often called mixing matrix or *dictionary* and source matrix) are unknown ($m \leq n < N$). We formulate conditions (SCA-conditions) under which we can recover $\mathbf{A}$ and $\mathbf{S}$ uniquely (up to scaling and permutation), such that $\mathbf{S}$ is *sparse* in the sense that each column of $\mathbf{S}$ has at least one zero element. We call this the *Sparse Component Analysis* problem (SCA). We present new algorithms for identification of the mixing matrix (under SCA-conditions), and for source recovery (under identifiability conditions). The methods are illustrated with examples showing good performance of the algorithms. Typical examples are EEG and fMRI data sets, in which the SCA algorithm allows us to detect some features of the brain signals. Special attention is given to the application of our method to the transposed system $\mathbf{X}^T = \mathbf{S}^T\mathbf{A}^T$ utilizing the sparseness of the mixing matrix $\mathbf{A}$ in appropriate situations. We note that the sparseness conditions could be obtained with some preprocessing methods and no independence conditions for the source signals are imposed (in contrast to Independent Component Analysis). We applied our method to fMRI data sets with dimension $(128 \times 128 \times 98)$ and to EEG data sets from a 256-channels EEG machine.

Key words: Sparse Component Analysis, Blind Signal Separation, clustering.

1 Introduction

Data mining techniques can be divided into the following classes [3]:

1. Predictive Modelling: where the goal is to predict a specific attribute (column or field) based on the other attributes in the data.

2. Clustering: also called segmentation, targets grouping the data records into subsets where items in each subset are more "similar" to each other than to items in other subsets.

3. Dependency Modelling: discovering the existence of arbitrary, possibly weak, multidimensional relations in data. Estimate some statistical properties of the found relations.

4. Data Summarization: targets finding interesting summaries of parts of the data. For example, similarity between a few attributes in a subset of the data.

5. Change and Deviation Detection: accounts for sequence information in data records. Most methods above do not explicitly model the sequence order of items in the data.

In this paper we consider the problem of linear representation or matrix factorization of a data set $\mathbf{X}$, given in the form of a $(m \times N)$-matrix:

$$\mathbf{X} = \mathbf{AS}, \quad \mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{S} \in \mathbb{R}^{n \times N}, \tag{1}$$

where n is the number of source signals, m is the number of observations and N is the number of samples. Such representations can be considered as a new class of data mining techniques (or a concrete subclass of the above described data mining technique 3). In (1) the unknown matrices $\mathbf{A}$ (dictionary) and $\mathbf{S}$ (signals) may have some specific properties, for instance:

1) the rows of $\mathbf{S}$ are as statistically independent as possible — this is the *Independent Component Analysis* (ICA) problem;

2) $\mathbf{S}$ contains as many zeros as possible — this is the sparse representation problem or *Sparse Component Analysis* (SCA) problem;

3) the elements of $\mathbf{X}$, $\mathbf{A}$ and $\mathbf{S}$ are nonnegative - this is *nonnegative matrix factorization* (NMF).

Such linear representations have several potential applications including decomposition of objects into "natural" components, learning the parts of the objects (e.g. learn from set of faces the parts a face consists of, i.e. eyes, nose, mouth, etc.), redundancy and dimensionality reduction, micro-array data mining, enhancement of images in nuclear medicine etc. (see [17], [10]).

There are many of papers devoted to ICA problems (see for instance [5, 15] and references therein) but mostly for the complete case ($m = n$). We refer to [26, 4, 29, 1, 25] and reference therein for some recent papers on SCA and overcomplete ICA ($m < n$).

A more general related problem is called *Blind Source Separation* (BSS) problem, in which we know *a priori* that a representation such as in equation (1) exists and the task is to recover the sources (and the mixing matrix) as accurately as possible. A fundamental property of the complete BSS problem (for $m = n$) is that such a recovery (under assumptions in 1 and non-Gaussianity of the sources) is possible up to permutation and scaling of the sources, which makes the BSS problem so attractive.

In this paper we consider SCA as a special model of BSS problem in the overcomplete case ($m < n$ i.e. more sources than sensors), where the additional information compensating the lack of sensors is the *sparseness* of the sources. The task of the SCA problem is to represent the given (observed) data $\mathbf{X}$ as in equation (1) such that the matrix $\mathbf{S}$ (sources) is sparse in sense that each column of $\mathbf{S}$ has at least one zero element. We present conditions on the data matrix $\mathbf{X}$ (*SCA-conditions on the data*), under which the representation in equation (1) is unique up to permutation and scaling of the sources.

The task of BSS problem is to estimate the unknown sources $\mathbf{S}$ (and the mixing matrix $\mathbf{A}$) using the available data matrix $\mathbf{X}$ only. We describe conditions (*identifiability conditions on the sources*) under which this is possible uniquely up to permutation and scaling of the sources, which is the usual condition in the complete BSS problems using ICA.

In the sequel, we present new algorithms for solving the BSS problem using sparseness: matrix identification algorithms and source recovery algorithm, which recovers sparse sources (in sense that each column of the source matrix $\mathbf{S}$ has at least one zero). When the sources are sufficiently sparse (see the conditions of Theorem 2) the matrix identification algorithm is even simpler. We used this simpler form for separation of mixtures of images. We present several computer simulation examples which illustrate our algorithms, as well as application of our method to real data: EEG data set obtained by a 256 channels EEG machine, and fMRI data set with dimension $128 \times 128 \times 98$. In all considered examples the results obtained by our SCA method are better (for the computer simulated examples) and comparable and advantages with respect to the ICA method.

2 Blind Source Separation using sparseness

In this section we present a method for solving the BSS problem if the following assumptions are satisfied:
A1) the mixing matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ has the property that any square $m \times m$ submatrix of it is nonsingular;
A2) each column of the source matrix $\mathbf{S}$ has at least one zero element.
A3) the sources are sufficiently rich represented in the following sense: for any index set of $n - m + 1$ elements $I = \{i_1, ..., i_{n-m+1}\} \subset \{1, ..., n\}$ there exist at least m column vectors of the matrix $\mathbf{S}$ such that each of them has zero elements in places with indexes in I and each $m - 1$ of them are linearly independent.

Columns of $\mathbf{X}$ for which A2) is not satisfied are called outliers. We can detect them in some cases and eliminate from the matrix $\mathbf{X}$, if the condition A3) is satisfied for a big number of columns of $\mathbf{S}$.

2.1 Matrix identification

We describe conditions in the sparse BSS problem under which we can identify the mixing matrix uniquely up to permutation and scaling of the columns. We give two types of such conditions. The first one corresponds to the least sparsest case in which such identification is possible. Further, we consider the most sparsest (nontrivial) case (for small number of samples) as in this case the algorithm is much simpler.

General case – full identifiability

Theorem 1. [12] (Identifiability conditions - general case) *Assume that the representation* $\mathbf{X} = \mathbf{AS}$ *is valid, the matrix* $\mathbf{A}$ *satisfies condition A1), the matrix* $\mathbf{S}$ *satisfies conditions A2) and A3) and only the matrix* $\mathbf{X}$ *is known. Then the mixing matrix* $\mathbf{A}$ *is identifiable uniquely up to permutation and scaling of the columns.*

The proof of this theorem is contained in [12] and gives the idea for the matrix identification algorithm.

Algorithm 1: identification of the mixing matrix

1) Cluster the columns of $\mathbf{X}$ in $\binom{n}{m-1}$ groups $\mathcal{H}_p, p = 1, ..., \binom{n}{m-1}$ such that the span of the elements of each group $\mathcal{H}_p$ produces one hyperplane and these hyperplanes are different.

2) Cluster the normal vectors to these hyperplanes in the smallest number of groups $G_j, j = 1, ..., n$ (which estimates the number of sources n) such that the normal vectors to the hyperplanes in each group G_j lie in a new hyperplane $\hat{H}_j$.

3) Calculate the normal vectors $\hat{\mathbf{a}}_j$ to each hyperplane $\hat{H}_j, j = 1, ..., n$ (the one-dimensional subspace spanned by $\hat{\mathbf{a}}_j$ is the intersection of all hyperplanes in G_j). The matrix $\hat{\mathbf{A}}$ with columns $\hat{\mathbf{a}}_j$ is an estimation of the mixing matrix (up to permutation and scaling of the columns).

Remark. The above algorithm works for data for which we know a priori that they lie on hyperplanes (or near to hyperplanes).

A very suitable algorithm for clustering data near hyperplanes is the k-plane clustering algorithm of Bradley - Mangasarian [2]. In our case the data points are supposed to lie on hyperplanes passing trough zero, so their algorithm is simplified and has the following form:

Algorithm 2: simplified algorithm of Bradley – Mangasarian

Start with random $w_1^0, ..., w_k^0 \in \mathbb{R}^n$ with $\|w_i^0\|_2 = 1, i = 1, ..., k$. Having $w_1^j, ..., w_k^j$ at iteration j with $\|w_i^j\|_2 = 1, i = 1, ..., k$, compute $w_1^{j+1}, ..., w_k^{j+1}$ by the following two steps:

(a) Cluster Assignment: Assign each point to closest plane P_l. For each $A_i, i = 1, ..., m$, determine $l(i)$ such that

$$|A_i w_{l(i)}^j| = \min_{1 \le j \le k} |A_i w_l^j|.$$

(b) Cluster Update: Find a plane P_l that minimizes the sum of the squares of distances to each point in cluster l. For $l = 1, ..., k$, let A_l be the $m(l) \times n$ matrix with rows corresponding to all A_i assigned to cluster l. Define $B(l) = [A(l)]^T A(l)$. Set w_l^{j+1} to be an eigenvector of $B(l)$ corresponding to the smallest eigenvalue of $B(l)$. Stop whenever there is a repeated overall assignment of points to cluster planes or a nondecrease in the overall objective function.

We applied this algorithm for real data sets of fMRI images and EEG recordings in the last two sections. We noticed that the algorithms stops very often in local minima and need several re-initializations until a reasonably good local (or global) minimum is found, measured by the nearness of the objective function to zero: the sum of the squared distances from the data points to the corresponding clustering hyperplanes should be near to zero.

Degenerate case – sparse instances

The following theorem is useful for identification of very sparse sources. Its proof can be found in [11].

Theorem 2. [11] (Identifiability conditions – locally very sparse representation) *Assume that (i) for each source $s_i := \mathbf{S}(i,.), i = 1, ..., n$ there are $k_i \ge 2$ time instances when all of the source signals are zero except s_i (so each source is uniquely present k_i times), and*

(ii) the set $\Big\{ j \in \{1, ..., N\} : \mathbf{X}(.,p) = c\mathbf{X}(.,j)$ for some $c \in \mathbb{R} \Big\}$, contains less than $\min_{1 \le i \le m} k_i$ elements for any $p \in \{1, ..., N\}$ for which $\mathbf{S}(.,p)$ has more that one nonzero element.

Then the matrix $\mathbf{A}$ is identifiable up to permutation and scaling.

Below we include an algorithm for identification of the mixing matrix in the case of Theorem 2.

Algorithm 3: identification of the mixing matrix in the very sparse case

1) Remove all zero columns of $\mathbf{X}$ (if any) and obtain a matrix $\mathbf{X}_1 \in \mathbb{R}^{m \times N_1}$.

2) Normalize the columns $\mathbf{x}_i, i = 1, \ldots, N_1$ of $\mathbf{X}_1 : \mathbf{y}_i = \mathbf{x}_i / \|\mathbf{x}_i\|$ and set $\varepsilon > 0$.

Multiply each column $\mathbf{y}_i$ by -1 if the first element of $\mathbf{y}_i$ is negative.

3) Cluster $\mathbf{y}_i, i = 1, ..., N_1$ in $n = 1$ groups $G_1, ..., G_{n+1}$ such that for any $i = 1, ..., n$, $\|\mathbf{x} - \mathbf{y}\| < \varepsilon, \forall \mathbf{x}, \mathbf{y} \in G_i$ and $\|\mathbf{x} - \mathbf{y}\| \ge \varepsilon$ for any $\mathbf{x}, \mathbf{y}$ belonging to different groups.

4) Chose any $\mathbf{y}_i \in G_i$ and put $\mathbf{a}_i = \mathbf{y}_i$. The matrix $\mathbf{A}$ with columns $\{\mathbf{a}_i\}_{i=1}^n$ is an estimation of the mixing matrix, up to permutation and scaling.

2.2 Identification of sources

Theorem 3. [12] (Uniqueness of sparse representation) *Let $\mathcal{H}$ be the set of all $\mathbf{x} \in \mathbb{R}^m$ such that the linear system $\mathbf{As} = \mathbf{x}, \mathbf{A} \in \mathbb{R}^{m \times n}$, has a solution with at least $n - m + 1$ zero components. If $\mathbf{A}$ fulfills A1), then there exists a subset $\mathcal{H}_0 \subset \mathcal{H}$ with measure zero with respect to $\mathcal{H}$, such that for every $\mathbf{x} \in \mathcal{H} \setminus \mathcal{H}_0$ this system has no other solution with this property.*

From Theorem 3 it follows that the sources are identifiable generically, i.e. up to a set with a measure zero, if they have level of sparseness grater than or equal to $n - m + 1$ (each column of $\mathbf{S}$ has at least $n - m + 1$ zeros) and the mixing matrix is known. Below we present an algorithm, based on the observation in Theorem 3.

Algorithm 4: source recovery algorithm

1. Identify the set of hyperplanes $\mathcal{H}$ produced by taking the linear hull of every subsets of the columns of $\mathbf{A}$ with $m - 1$ elements;

2. Repeat for $k = 1$ to N:

2.1. Identify the space $H \in \mathcal{H}$ containing $\mathbf{x}_k := \mathbf{X}(:, k)$, or, in practical situation with presence of noise, identify the one to which the distance from $\mathbf{x}_i$ is minimal and project $\mathbf{x}_k$ onto H to $\tilde{\mathbf{x}}_k$;

2.2. if H is produced by the linear hull of column vectors $\mathbf{a}_{k_1}, ..., \mathbf{a}_{k_{m-1}}$, then find coefficients $\lambda_{k,j}$ such that

$$\tilde{\mathbf{x}}_k = \sum_{j=1}^{m-1} \lambda_{k,j} \mathbf{a}_{k_j}.$$

These coefficients are uniquely determined if $\tilde{\mathbf{x}}_k$ doesn't belong to the set $\mathcal{H}_0$ with measure zero with respect to $\mathcal{H}$ (see Theorem 3);

2.3. Construct the solution $\mathbf{s}_k = \mathbf{S}(:, k)$: it contains $\lambda_{k,j}$ in the place k_j for $j = 1, ..., m - 1$, the rest of the components are zero.

3 Sparse Component Analysis

In this section we describe sufficient conditions for the existence of solutions to the SCA problem. Note that the conditions are formulated only in terms of the data matrix $\mathbf{X}$. The proof of the following theorem can be found in [12].

Theorem 4. [12] (SCA conditions) *Assume that $m \leq n \leq N$ and the matrix $\mathbf{X} \in \mathbb{R}^{m \times N}$ satisfies the following conditions:*

(i) the columns of $\mathbf{X}$ lie in the union $\mathcal{H}$ of $\binom{n}{m-1}$ different hyperplanes, each column lies in only one such hyperplane, each hyperplane contains at least m columns of $\mathbf{X}$ such that each $m - 1$ of them are linearly independent.

(ii) for each $i \in \{1, ..., n\}$ *there exist* $p = \binom{n-1}{m-2}$ *different hyperplanes* $\{H_{i,j}\}_{j=1}^{p}$ *in* $\mathcal{H}$ *such that their intersection* $L_i = \cap_{k=1}^{p} H_{i,j}$ *is one dimensional subspace.*

(iii) any m *different* L_i *span the whole* $\mathbb{R}^m$*.*

Then the matrix $\mathbf{X}$ *is representable uniquely (up to permutation and scaling of the columns of* $\mathbf{A}$ *and* $\mathbf{S}$*) in the form* $\mathbf{X} = \mathbf{AS}$*, where the matrices* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *and* $\mathbf{S} \in \mathbb{R}^{n \times N}$ *satisfy the conditions A1), A2), and A3) respectively.*

4 Overdetermined Blind Source Separation

In this section we assume that $m > n$ and the identifiability conditions for the transposed matrix $\mathbf{A}^T$ are satisfied. So we have the model:

$$\mathbf{X}^T = \mathbf{S}^T \mathbf{A}^T, \tag{2}$$

but in order to apply Theorem 1 we select n rows of the matrices $\mathbf{X}^T$ and $\mathbf{S}^T$ (usually the first n, assuming that they (for $\mathbf{S}^T$) are linearly independent: this is true with "probability one", i.e. the matrices without this property form a set with measure zero). Denoting $\mathbf{X}_n = \mathbf{X}(:, 1:n)$ and $\mathbf{S}_n = \mathbf{S}(:, 1:n)$, we have

$$\mathbf{X}_n^T = \mathbf{S}_n^T \mathbf{A}^T. \tag{3}$$

By some of the matrix identification algorithms we identify firstly the matrix $\mathbf{S}_n^T$ and then we identify the matrix $\mathbf{A}$: $\mathbf{A} = \mathbf{X}_n \mathbf{S}_n^{-1}$. Now we recover the full matrix $\mathbf{S}$ from (2) by $\mathbf{S} = \mathbf{A}^+ \mathbf{X}$, where $\mathbf{A}^+$ means the Moore-Penrose pseudo-inverse of $\mathbf{A}$.

5 Computer simulation examples

5.1 Overdetermined Blind Source Separation – very sparse case

We consider the overdetermined mixture of two artificially created *non-independent* and *non-sparse* sources with 10 samples – see Figure 1. The mixing matrix and the estimated matrix with the overcomplete blind source separation scheme (see section 4) are respectively

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 2 & 3 \\ 2 & 0 \\ 1 & 1 \\ 1 & 5 \\ 0 & -1 \\ 1 & 0 \end{pmatrix} \qquad \text{and} \qquad \hat{\mathbf{A}} = \begin{pmatrix} 0 & 1.2 \\ 7.3 & 3.6 \\ 7.3 & 0 \\ 3.6 & 1.2 \\ 3.6 & 6.1 \\ 0 & -1.2 \\ 3.6 & 0 \end{pmatrix}.$$

The mixtures and estimated sources are shown in Figure 1. In this case we applied Algorithm 3 for identification of the matrix $\mathbf{S}_2^T$ (the transposed of

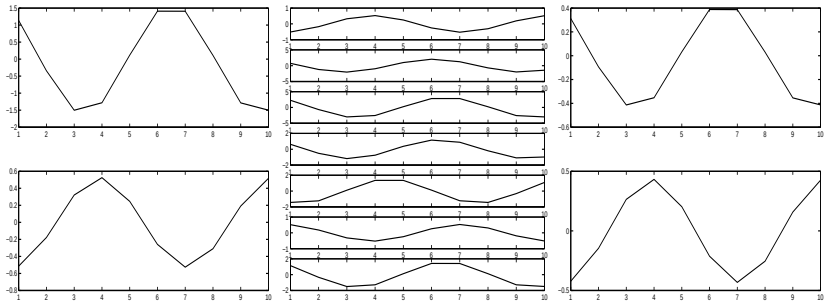


Fig. 1. Example 1. Left:Artificially created *non-independent* and *non-sparse* source signals. Middle: Their mixtures with matrix $\mathbf{A}$. Right: Recovered source signals. The signal-to-noise ratio between the original sources and the recoveries is very high with 319 and 319 dB after permutation and normalization.

the first two rows of the source matrix $\mathbf{S}$, see (3)). After normalization of each row of $\hat{\mathbf{A}}$ we obtain the original matrix $\mathbf{A}$, which confirms the perfect reconstruction of the sources. The transposed matrix $\mathbf{A}^T$ (considered here as a new source matrix) satisfies the conditions of Theorem 2 and this is the reason for the perfect reconstruction of the sources.

5.2 Overdetermined Blind Source Separation - Sparse Case

Now let us consider the overdetermined mixture of 3 artificially created *non-independent* and *non-sparse* sources with only 10 samples (in fact only *three* are needed, as in the previous example only two were needed) — see Figure 2 (left).

The mixing matrix and the estimated matrix with the overcomplete blind source separation scheme (see section 4) are respectively

$$\mathbf{A} = \begin{pmatrix} 0.5287 & 0.5913 & 0 \\ 0.2193 & -0.6436 & 0 \\ -0.9219 & 0.3803 & 0 \\ -2.1707 & 0 & 0.7310 \\ -0.0592 & 0 & 0.5779 \\ -1.0106 & 0 & 0.0403 \\ 0 & 0.0000 & 0.6771 \\ 0 & -0.3179 & 0.5689 \\ 0 & 1.0950 & -0.2556 \end{pmatrix}, \hat{\mathbf{A}} = \begin{pmatrix} 0.0000 & 0.8631 & 0.7667 \\ -0.0000 & -0.9395 & 0.3180 \\ -0.0000 & 0.5552 & -1.3368 \\ 1.0972 & -0.0000 & -3.1476 \\ 0.8674 & -0.0000 & -0.0858 \\ 0.0605 & 0.0000 & -1.4655 \\ 1.0164 & 0.0001 & -0.0000 \\ 0.8540 & -0.4640 & -0.0000 \\ -0.3837 & 1.5984 & 0.0000 \end{pmatrix}.$$

Now we apply Algorithm 1 – note that only 9 samples are required by the identifiability theorem – Theorem 1 (due to condition A3)), and $\mathbf{A}^T$ has precisely 9 rows. The mixtures are shown in Figure 3, along with a scatter plot for a visualization of the matrix detection in this transposed case with the *very* low sample number of only 9, which is sufficient for a perfect recovery

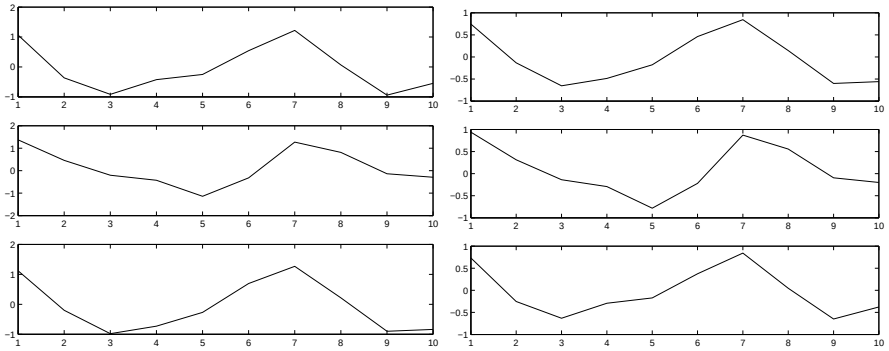


Fig. 2. Example 2. Left: Artificially created *non-independent* and *non-sparse* source signals. Right: Recovered source signals. The signal-to-noise ratio between the original sources and the recoveries is very high with 308, 293 and 307 dB after permutation and normalization.

of (transposed) mixing matrix and the original sources (estimated sources are shown in Fig. 2 right).

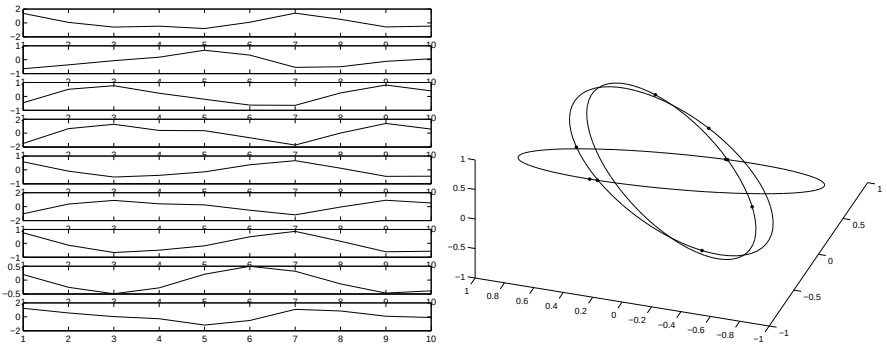


Fig. 3. Example 2. Left: mixed signals $\mathbf{X}$ (observed sources). Right: Scatterplot of the new 'observed sources' $\mathbf{X}_3^T$ (after transposition of $\mathbf{X}_3$ – the first 3 data samples) together with the hyperplanes on which they lie, indicated by their intersections with the unit sphere (circles).

5.3 Complete case

In this example for the complete case ($m = n$) of instantaneous mixtures, we demonstrate the effectiveness of our algorithm for identification of the mixing matrix in the case considered in Theorem 2. We mixed 3 images of

landscapes (shown in Fig. 4) with a 3-dimensional randomly generated matrix $\mathbf{A}$ (det$\mathbf{A} = 0.0016$). We transformed these three mixtures (shown in Fig. 5) by two dimensional discrete Haar wavelet transform and took only the 10-th row (160 points) of the obtained diagonal coefficients $cD\mathbf{X}$. As a result, since this transform is linear, the corresponding diagonal wavelet coefficients $cD\mathbf{S}$ of the source matrix $\mathbf{S}$ represented by the source images (as well as the horizontal and vertical ones) become very sparse (see Fig. 7) and they satisfy the conditions of Theorem 2. Using only one row (the 10-th or any other, with 160 points) of $cD\mathbf{X}$ appears to be enough to estimate very precisely the mixing matrix, and therefore, the original images. The estimated images are shown in Fig. 6.

Fig. 4. Original images

Fig. 5. Mixed (observed) images

Fig. 6. Estimated normalized images using the estimated matrix. The signal-to-noise ratios with the sources from Figure 1 are 232, 239 and 228 dB respectively.

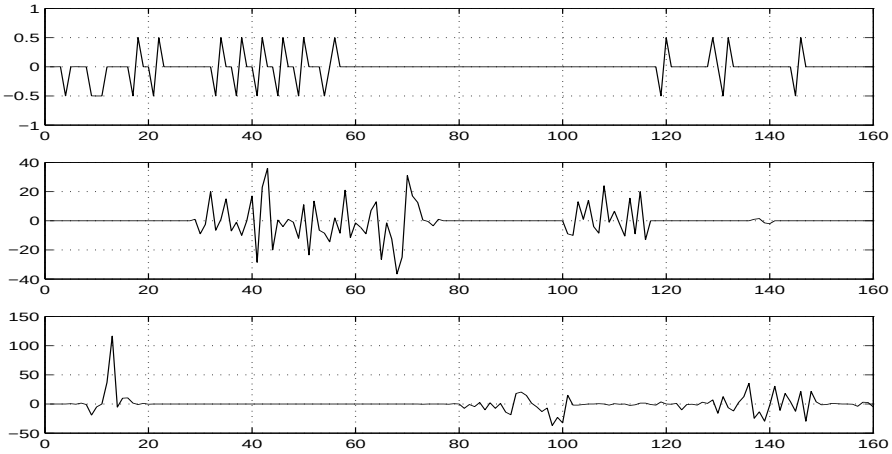


Fig. 7. Diagonal wavelet coefficients of the original images (displaying only the 10-th row of each of the three (120×160) matrixes). They satisfy the conditions of Theorem 1 and this is the reason for the perfect reconstruction of the original images, since our algorithm uses only the tenth row of each of the mixed images.

5.4 Underdetermined case

We consider a mixture of 7 artificially created sources (see Fig. 9 left) – sparsified randomly generated signals with at least 5 zeros in each column – with a randomly generated mixing matrix with dimension 3×7.

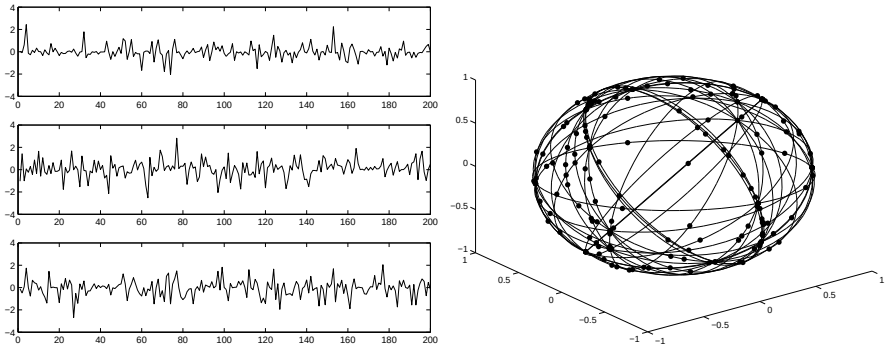


Fig. 8. Mixed signals (left) and normalized scatter plot (density) of the mixtures (right) together with the 21 data set hyperplanes, visualized by their intersection with the unit sphere in $\mathbb{R}^3$.

Figure 8 gives the mixed signals together with a normalized scatterplot of the mixtures – the data lies in $21 = \binom{7}{2}$ hyperplanes.

Applying the underdetermined matrix recovery algorithm (Algorithm 1) to the mixtures gives the recovered mixing matrix exactly, up to permutation and scaling. Applying the source recovery algorithm (Algorithm 4) we recover the source signals up to permutation and scaling (see Fig. 9, middle). This figure (right) shows also that the recovery by l_1-norm minimization (known as Basis Pursuit method of S. Chen, D. Donoho and M. Saunders [7]) does not perform well, even if the mixing matrix is perfectly known.

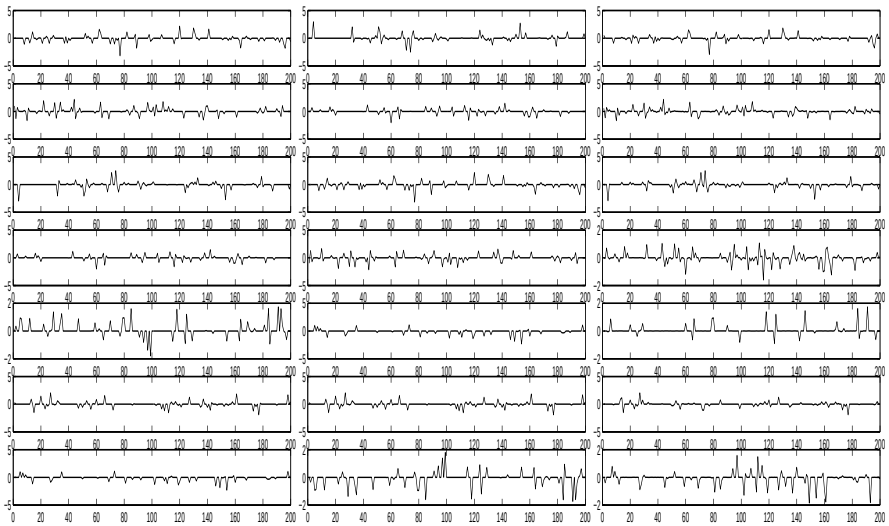


Fig. 9. The original source signals are shown in the left column. The middle column gives the recovered source signals — the signal-to-noise ratio between the original sources and the recoveries is very high (above 278 dB after permutation and normalization). Note that only 200 samples are enough for excellent separation. The right column shows the recovered source signals using l_1-norm minimization and known mixing matrix. Simple comparison confirms that the recovered signals are far from the original ones – the signal-to-noise ratio is only around 4 dB.

6 Extraction of auditory evoked potentials from EEG contaminated by eye movements by Sparse Component Analysis

6.1 Introduction

Ocular artifact contamination is very common in electroencephalographic (EEG) recordings. The electro-oculographic (EOG) signals are generated by the horizontal movement of the eyes, which act as charged electric dipoles with the positive poles at the cornea and the negative poles at the retina. These electric charges of the movement are picked up by frontal EEG electrodes. The EOG contamination is normally dealt with by instructing the subjects not to blink and not to move the eyes during an EEG experiment, as well as by trying to reject the affected data using voltage threshold criteria. Both of these measures leave a lot to be desired, because cognitive commands to subjects may introduce additional complexity, while at the same time very slow eye movements are difficult to identify only by voltage thresholding because their amplitudes may be comparable to those of the underlying electroencephalogram. Recent studies have proposed artifact removal procedures based on estimation of correction coefficients [8] and independent component analysis [13], [19], [14], [18], [20], etc. The goal of the present section is to demonstrate that the new Sparse Component Analysis (SCA) method extracts efficiently for further usage the underlying evoked auditory potentials masked by strong eye movements.

6.2 Methods

The electric potentials on the surface of the scalp of human subjects were measured with a geodesic sensor net using a 256-channel electroencephalographic (EEG) system (Electrical Geodesics Inc., Eugene, Oregon, USA). An on-screen pattern image was presented for scanning 20 times. During each presentation the subject had to scan 4 lines - two horizontal and two vertical. A button was pressed by the subject immediately before a line scan and another button - signaling that the line scan was completed. A 1000 Hz, 100ms, 100 dB sound accompanied the pattern image each time after the start button was activated. An eye tracking device (EyeGaze , LC Technologies, Inc.) was used for precision recording and control of all eye movements during the EEG experiments, scanning the subjects' screen gaze coordinates 60 times per second.

After aligning 20 single epochs of each eye movement type (horizontal left-to-right, diagonal down-left, horizontal right-to-left, diagonal up-left) with their corresponding sound stimulus onset times, EEG data was segmented into trials of 500ms lengths and averaged separately for each line type. We then preprocessed the segmented contaminated data by reducing its dimensionality from 256 to 4 using principal component analysis (PCA). A justification for

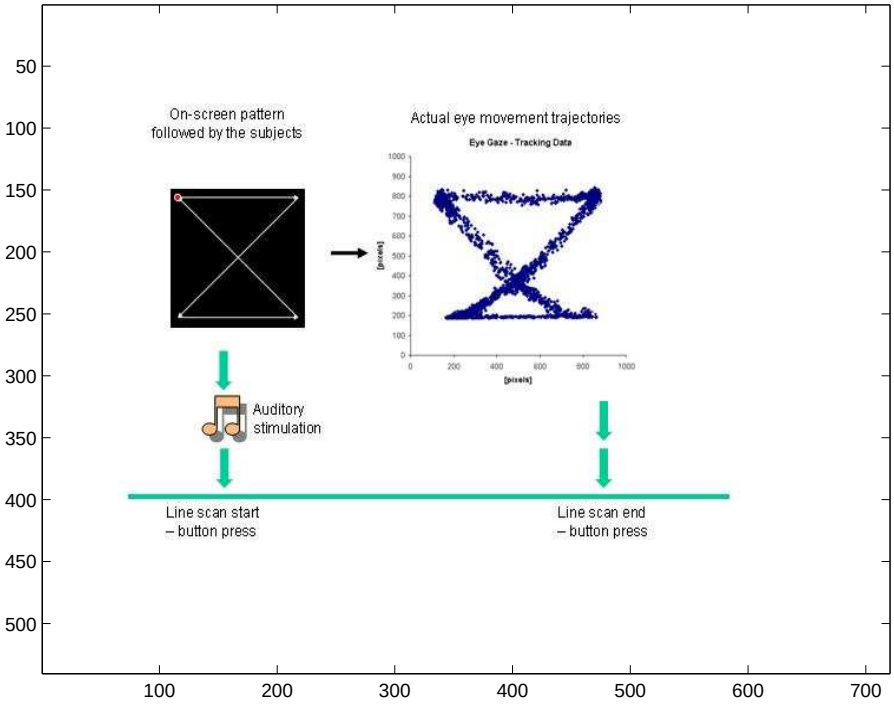


Fig. 10. Experimental design to extract auditory evoked potentials from high-density EEG data contaminated by eye movements. Ocular artifacts were controlled by an eye tracking device.

such a reduction is shown in Fig. 11 which shows that the subspace spanned by the principal components corresponding to the biggest 4 singular values contain most of the information of the data. The new Sparse Component Analysis method was applied on this new data set and its performance was compared to basic ICA algorithms: Fast ICA algorithm and JADE algorithm (see [15] for reference about ICA methods and algorithms).

6.3 Results

In this experiment we applied Algorithm 2 for matrix identification (using several re-initializations, until obtaining satisfactory local (or global) minimum of the cost function: the sum of the squared distances from the data points to the corresponding clustering hyperplanes should be small. For source recovery we apply either Algorithm 4, or inversion of the estimated matrix: the results are similar, and as in the the inversion matrix method the resulting signals are slightly more smooth.

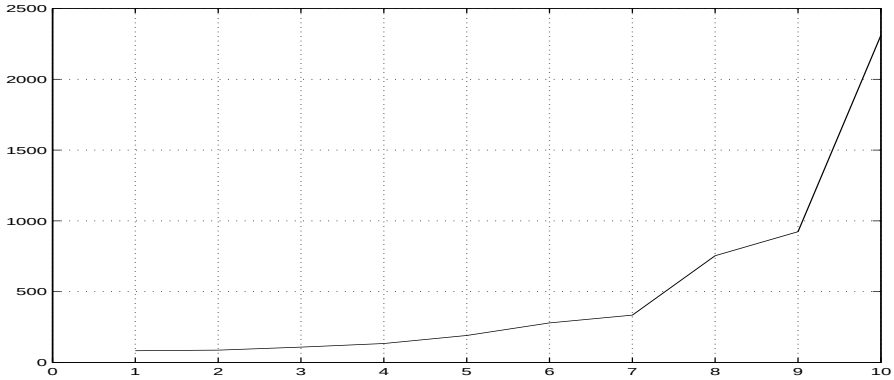


Fig. 11. The biggest 10 singular values of the data matrix from 256 channels EEG machine

The component related to the evoked auditory N1 potential [23] with a peak amplitude at 76-124ms [24] was identified (with some differences) by all applied algorithms. However, SCA algorithm gives the best result (Fig. 12 right, 4-th component) which correspond to the reality of the experiment, i.e. the auditory stimulus was silent after 150 ms. The Fast ICA and JADE algorithms (Fig. 13) show nonzero activity in the auditory component after 150 ms, which is false. The eye movements, however, were strongly mixed and masked, so that the various algorithms presented different performance capabilities. SCA's component 1 (Fig. 12 right) corresponded to the steady and continuous horizontal eye movement of the subject from the left side of the image to right side. The initial plateau (0–170 ms) was due to the subjective delay before the subject was able to start the actual movement. FastICA and JADE (Fig. 13, 3-rd left and 1-st right components respectively) were unable to reveal fully the underlying continuous potentials resulting from the eye movement. SCA component 3 was slightly different at 250-300 ms, but overall similar to component 4, which could have indicated that the real number of strong sources was 3 and this component was redundant. However, if that was not the case, then both this component, as well as SCA component 2 were either of eye movement origin and had been caused by acceleration jolt in response to the sound startle effect, or were related to the button press motor response potentials in cortex.

In order to verify or reject the hypothesis that the real number of strong sources was 3 (and SCA component 3 in Fig. 12, right, was redundant), we performed similar processing with just 3 input signals extracted by PCA signal reduction. The results are shown in Fig. 14 (right) and Fig. 15. Again, the auditory response was mixed with the eye movement potentials in the input data (Fig. 14 left) and all three algorithms were able to obtain the N1 evoked potential - SCA (Fig. 14 right, 3-rd component), FastICA and JADE (Fig.

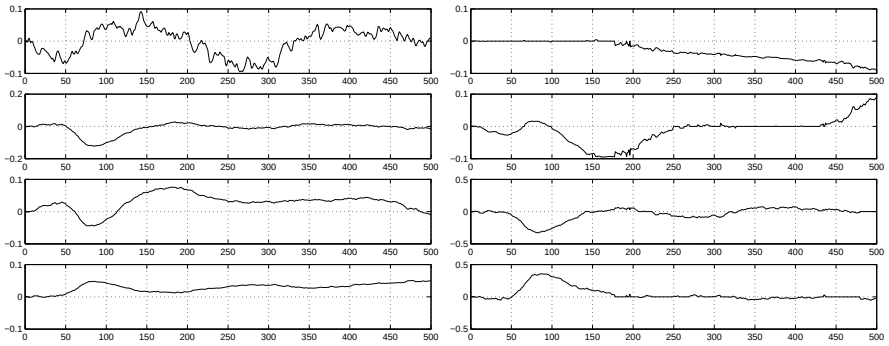


Fig. 12. Left: Input EEG data with dimensionality reduced from 256 channels to 4 principal components. This data was not sufficiently separated and still contained mixed information about the original cortical and eye dipole sources. Right: Sparse Component Analysis (SCA) results.

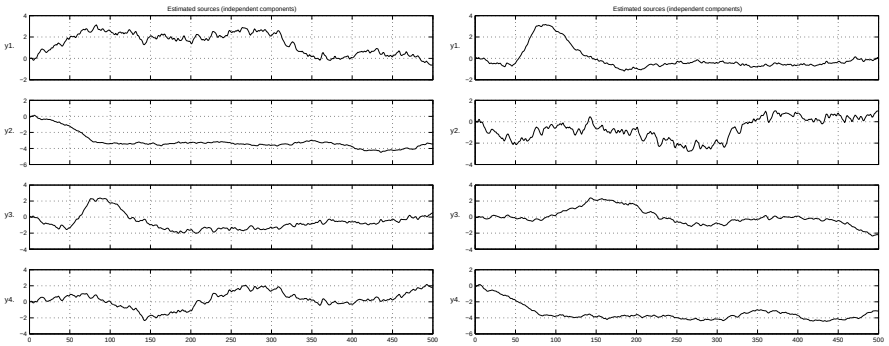


Fig. 13. Left: FastICA results. Right: JADE results.

15, 2-nd and 1-st components respectively), as those found by SCA is minimally deviated from zero in the period 150-500 ms. However, the steady eye movement ramp was most difficult to extract by the ICA methods FastICA (Fig. 15), while SCA (Fig. 14 right, 3-rd component) revealed again a clear basic trend potential without overlapping peaks. SCA component 2 (Fig. 14 right) was represented in a varying degree also by the ICA algorithms.

Our SCA method exhibited a best fit for the hypothesis with 3 sources of electrical potentials in the mixed auditory and eye movement data. Nevertheless, additional experiments may be needed to better reveal the rather complex structure of the eye movement signal.

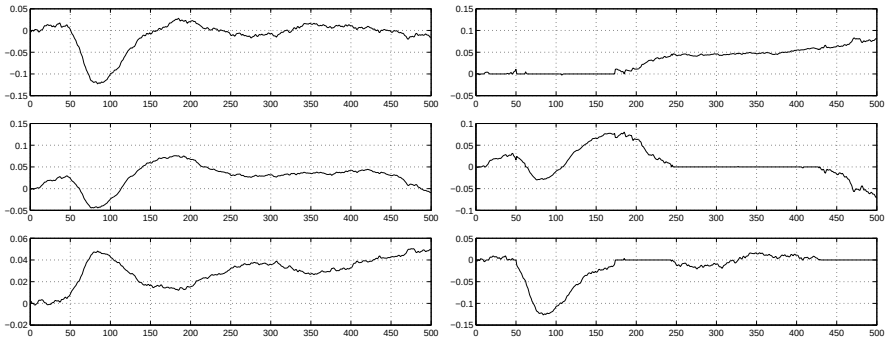


Fig. 14. Left: Input EEG data with dimensionality reduced from 256 channels to 4 principal components. This data was not sufficiently separated and still contained mixed information about the original cortical and eye dipole sources. Right: Sparse Component Analysis (SCA) results.

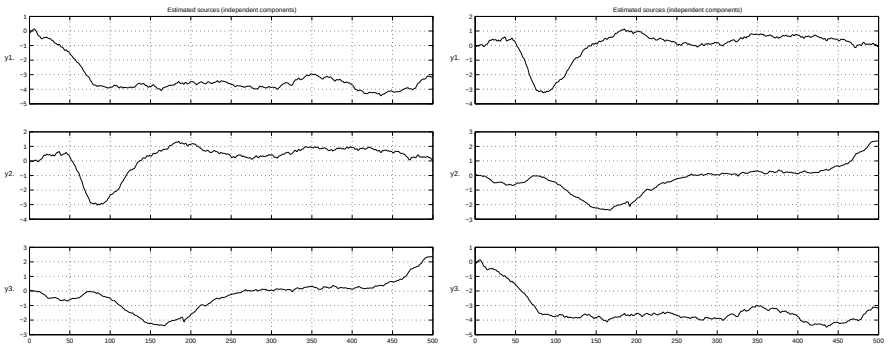


Fig. 15. Left: FastICA results. Right: JADE results.

7 Applications of Sparse Component Analysis to fMRI data

7.1 SCA applied to fMRI toy data

We simulated a low-dimensional example of fMRI data analysis. The typical setup of fMRI experiments is the following: NMR brain imaging techniques are used to record brain activity data over a certain span of time, during which the subject is asked to perform some kind of task (e.g. 5 seconds of activity in the motor cortex followed by 5 seconds of activity in the visual cortex; this iterative procedure is often called *block diagram*). The brain recordings show areas of high and of low brain activity (using the *BOLD effect*). Analysis is performed on the 2d-image slices recorded at the discrete time steps. General linear model (GLM) approaches or ICA-based fMRI analysis then decompose

this data set into a certain set of *component maps* i.e. sets of (hopefully independent) images that are active at certain time steps corresponding to the block diagram.

In the following we simulate a low-dimensional example of such brain activity recordings. For this we mix three 'source component maps' (Fig. 16) linearly to three mixture images and add some noise.

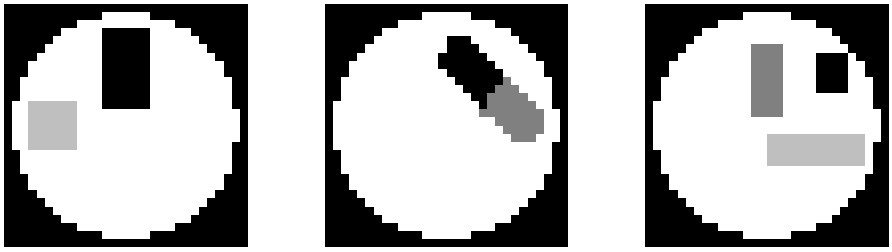


Fig. 16. Example: artificial *non-independent* and *non-sparse* source signals.

These mixtures represent our recordings at three different time steps. From the recordings we want to recover the original components or component maps. We want to use an unsupervised approach (not GLM, which requires additional knowledge of the mixing system) but with a different contrast than ICA. We believe that the assumption of *independence* of the component maps does not hold in a lot of situations, so we replace this assumption by *sparseness* of the maps, meaning that at a certain voxel, not all maps are allowed to be active (in the case of as many mixtures as sources).

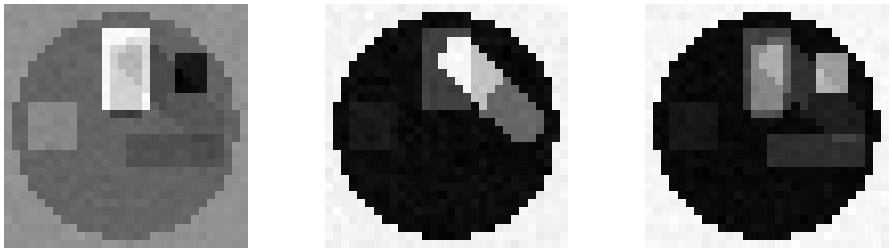
We consider a mixture of 3 artificially created *non-independent* source images of size 30×30 — see Figure 16 — with the (normalized) mixing matrix

$$\mathbf{A} = \begin{pmatrix} -0.9069 & 0.1577 & 0.4726 \\ -0.2737 & -0.9564 & 0.0225 \\ -0.3204 & -0.2458 & -0.8810 \end{pmatrix}$$

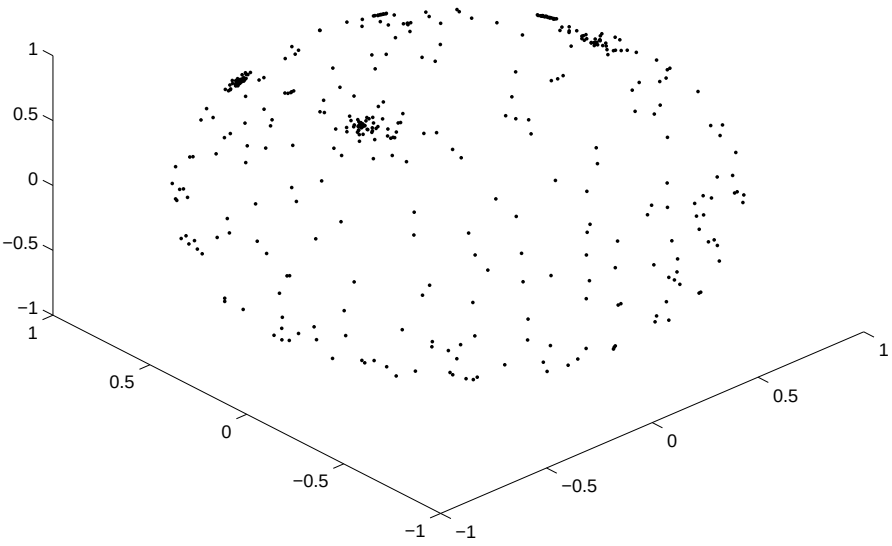
and 4% of additive white noise. The mixtures are shown in Figure 17 together with their scatterplot after normalization to unit length.

Note that due to the circular 'brain region', we have to preprocess the data ('sparsification') by removing the non-brain voxels from the boundary. Then, we apply the matrix identification algorithm (Algorithm 1). This gives the recovered matrix (after normalization)

$$\hat{\mathbf{A}} = \begin{pmatrix} 0.9110 & 0.1660 & 0.4693 \\ 0.2823 & -0.9541 & 0.0135 \\ 0.3007 & -0.2494 & -0.8829 \end{pmatrix}$$



(a)



(b)

Fig. 17. Example: mixed signals with 4% additive noise (a), and scatterplot after normalization to unit length (b).

with low crosstalking error 0.12 and the recovered sources $\hat{\mathbf{S}}$ shown in Figure 18, with high signal-to-noise ratio of 28, 27 and 27 dB with respect to the original sources (after permutation and normalization).

This can be enhanced by applying a denoising algorithm to each image. Figure 19 shows the application of local PCA denoising with an MDL-parameter estimation criterion, which gives SNRs of 32, 31 and 29 dB, so a mean enhancement of around 4 dB has been achieved.

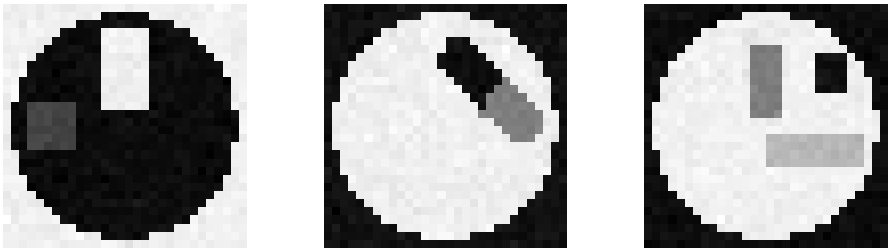


Fig. 18. Example: recovered source signals. The signal-to-noise ratio between the original sources (figure 16) and the recoveries is high with 28, 27 and 27 dB after permutation and normalization.

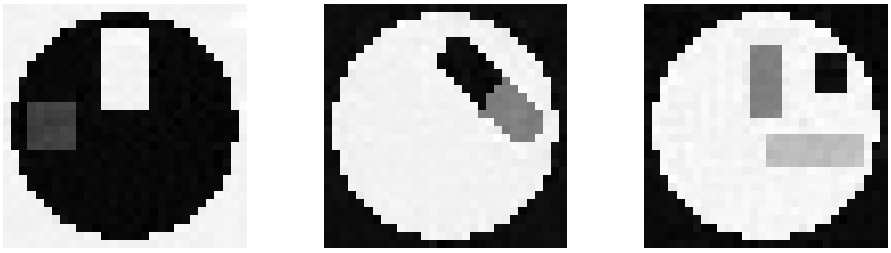


Fig. 19. Example: recovered denoised source signals. Now the SNR is even higher than in figure 18 (32, 31 and 29 dB after permutation and normalization).

Note that if we apply ICA to the previous example (after sparsification as above — without sparsification ICA performs even worse), the algorithm cannot recover the mixing matrix

$$\bar{\mathbf{A}} = \begin{pmatrix} 0.6319 & -0.3212 & 0.8094 \\ -0.0080 & -0.8108 & -0.3138 \\ -0.7750 & -0.4893 & 0.4964 \end{pmatrix}$$

and has a very high crosstalking error of 4.7 with respect to $\mathbf{A}$. Figure 20 shows the poorly recovered sources; the SNRs with respect to the sources are only 3.3, 13 and 12 dB respectively. The reason for ICA not being able to recover the sources simply lies in the fact that they were not chosen to be independent.

7.2 SCA applied to real fMRI data

We now analyze the performance of SCA when applied to real fMRI measurements. fMRI data were recorded from six subjects (3 female, 3 male, age 20–37) performing a visual task. In five subjects, five slices with 100 images (TR/TE = 3000/60 msec) were acquired with five periods of rest and five

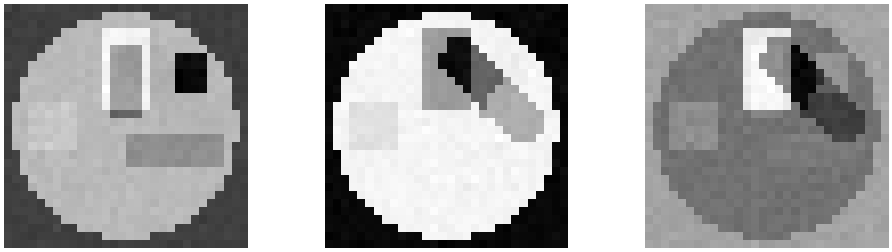


Fig. 20. Example: poorly recovered source signals using ICA. The signal-to-noise ratio between the original sources (figure 16) and the recoveries is very low with 3.3, 13 and 12 dB after permutation and normalization.

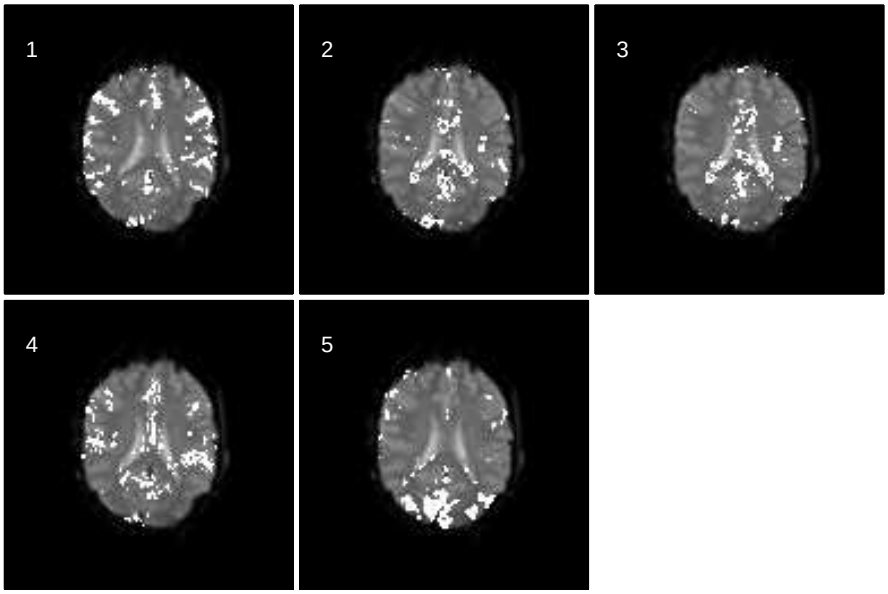
photic simulation periods with rest. Simulation and rest periods comprised 10 repetitions each, i.e. 30s. Resolution was $3 \times 3 \times 4$ mm. The slices were oriented parallel to the calcarine fissure. Photic stimulation was performed using an 8 Hz alternating checkerboard stimulus with a central fixation point and a dark background with a central fixation point during the control periods [27]. The first scans were discarded for remaining saturation effects. Motion artifacts were compensated by automatic image alignment (AIR, [28]).

Blind Signal Separation, mainly based on ICA, nowadays is a quite common tool in fMRI analysis (see for example [21], [22]). Here, we analyze the fMRI data set using as a separation criterion a spatial decomposition of fMRI data images to sparse component maps. Such an approach we consider as very reasonable and advantageous when the stimuli are sparse and dependent, and therefore the ICA methods couldn't give good results. Due to the availability of fMRI data, it appears that the results of our SCA method and ICA method give similar results, which itself we consider as a surprising fact. Here we use again Algorithm 2 for matrix identification and Algorithm 4 or matrix inversion of the estimated matrix, for estimation of the sources.

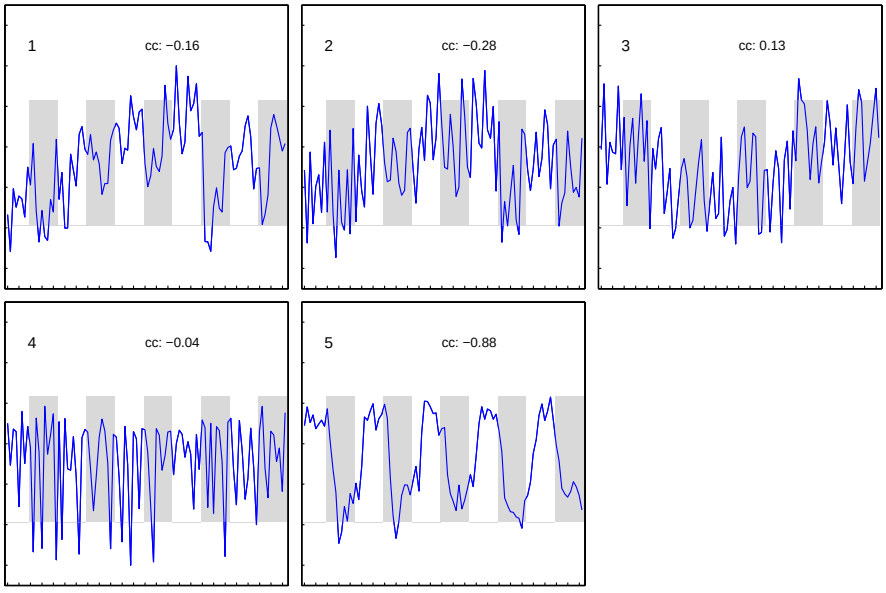
Figure 21 shows the performance of SCA method; see figure caption for interpretation. Using only the first 5 principal components, SCA could recover the stimulus component as well as detect additional components. It performs equally well as fastICA, Figure 22, which is interesting in itself: apparently the two different criteria, sparseness and independence, lead to similar results in this setting. This can be partially explained by noting that all components, mainly the stimulus component, have high kurtoses i.e. strongly peaked densities.

8 Conclusion

We rigorously defined the SCA and BSS problems of sparse signals and presented sufficient conditions for their solution. We presented four algorithms

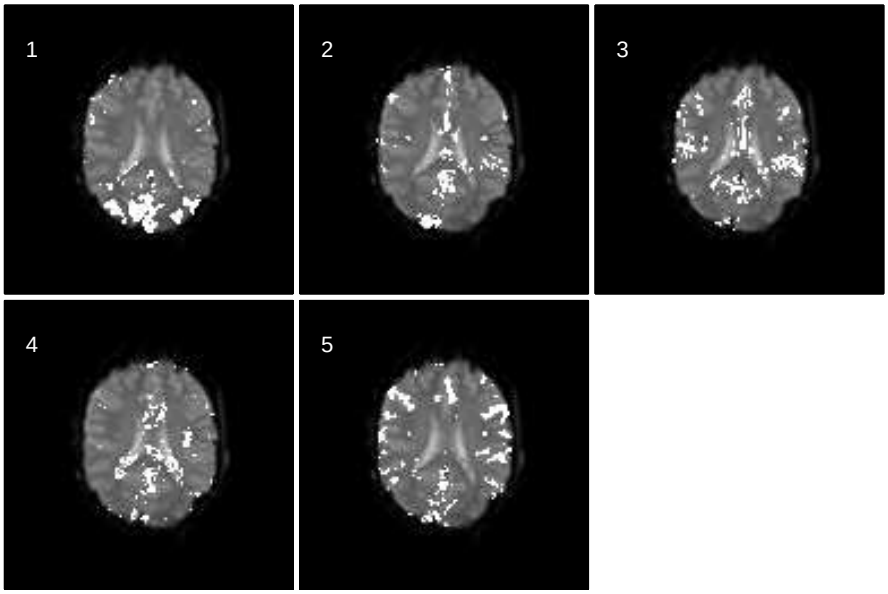


(a) component maps

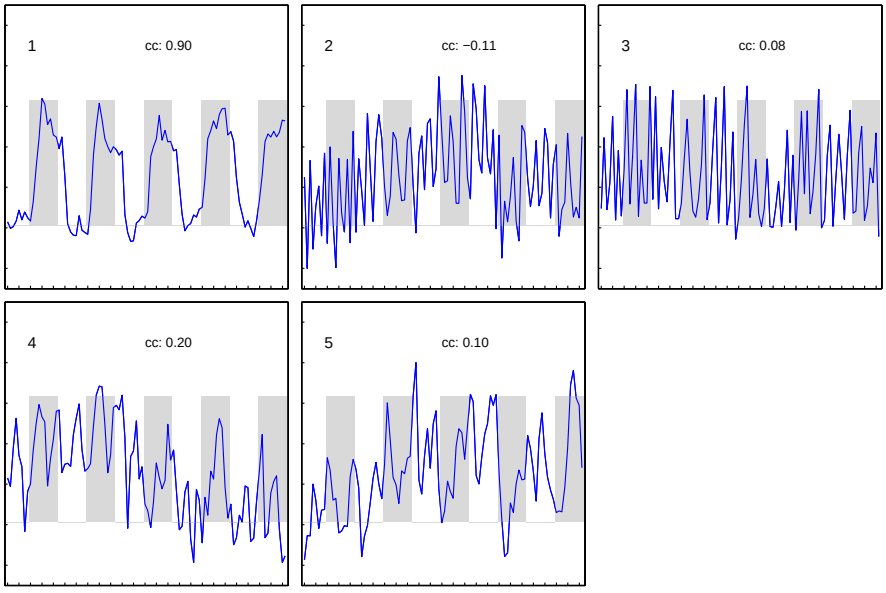


(b) time courses

Fig. 21. SCA fMRI analysis. The data was reduced to the first 5 principal components. (a) shows the recovered component maps (white points indicate values stronger than 3 standard deviations), and (b) their time courses. The stimulus component is given in component 5 (indicated by the high crosscorrelation $cc = -0.86$ with the stimulus time course, delayed by roughly 2 seconds due to the BOLD effect), which is strongly active in the visual cortex as expected.



(a) component maps



(b) time courses

Fig. 22. FastICA result during fMRI analysis of the same data set as in figure 21. The stimulus component is given in component 1 with high stimulus cross-correlation $cc = 0.90$.

applicable to SCA: one for source recovery and three ones for identification of the mixing matrix – for the sparse and the very sparse cases and one based on a simplified Bradley-Mangasarian's k-plane clustering algorithm. We presented several experiments for confirmation of our methods, including applications in fMRI and EEG data sets.

Although it is a standard practice to cut those evoked-potentials in EEG data which are contaminated by eye movement artifacts, we have demonstrated that stimulus-related responses could be recovered successfully and even better by the Sparse Component Analysis method. In addition, SCA has revealed a complex hidden structure of the dynamically accelerating eye movement signal, which could become a future basis for a new instrument to measure objectively individual psychological characteristics of a human subject in startle reflex-type experiments, exploiting sparseness of the signals rather than independence. We have also shown that our new method is a useful tool in separating the functional EEG components more efficiently in signal hyperspace than independent component analysis. Very promising are the results with real fMRI data images, which show that revealing the brain responses of sparse (and may be dependent) stimuli could be more successful by SCA than by ICA.

Acknowledgements

The authors would like to thank Dr. Dorothee Auer from the Max Planck Institute of Psychiatry in Munich, Germany, for providing the fMRI data, and Oliver Lange from the Department of Clinical Radiology, Ludwig-Maximilian University, Munich, Germany, for data preprocessing of fMRI data and visualization.

References

1. P. Bofill and M. Zibulevsky, " Underdetermined Blind Source Separation using Sparse Representation", Signal Processing, vol. 81, no. 11, pp. 2353-2362, 2001.
2. P.S. Bradley and O. L. Mangasarian, "k-Plane Clustering", *J. Global optim.*, 16, (2000), no.1, 23-32.
3. P.S. Bradley, U. M. Fayyad and O. L. Mangasarian, "Mathematical progrmming for data mining: formulations and challenges", *INFORMS J. Computing*, **11** (1999), no. 3, 217–238.
4. A.M. Bronstein, M.M. Bronstein, M. Zibulevsky, Y. Y. Zeevi, " Blind Separation of Reflections using Sparse ICA", in Proc. Int. Conf. ICA2003, Nara, Japan, pp.227-232.
5. A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*. John Wiley, Chichester, 2002.

6. A. Cichocki, S. Amari, K. Siwek, "ICALAB for signal processing package", http://www.bsp.brain.riken.jp
7. S. Chen, D. Donoho and M. Saunders, " Atomic decomposition by basis pursuit", *SIAM J. Sci. Comput.*, Vol. 20, no. 1, pp. 33–61, 1998.
8. Croft RJ, Barry RJ., "Removal of ocular artifact from the EEG: a review", Neurophysiol. Clin. 2000, Feb 30, pp. 5-19.
9. D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization", *Proc. Nat. Acad. Sci.*, vol.100, no.5, pp. 2197–2202, 2003.
10. D. Donoho and V. Stodden, "When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts?", Neural Information Processing Systems (NIPS) 2003 Conference, http://books.nips.cc
11. P. G. Georgiev and A. Cichocki, "Sparse component analysis of overcomplete mixtures by improved basis pursuit method", accepted in 2004 IEEE International Symposium on Circuits and Systems (ISCAS 2004).
12. P.G. Georgiev, F. Theis and A. Cichocki, "Blind Source Separation and Sparse Component Analysis of overcomplete mixtures", accepted in ICASSP 2004 (International Conference on Acoustics and Statistical Signal Processing).
13. Georgiev P., Cichocki A., Bakardjian H., "Optimization Techniques for Independent Component Analysis with Applications to EEG Data", In: Pardalos et al, editors, Quantitative Neuroscience: Models, Algorithms, Diagnostics, and Therapeutic Applications, Kluwer Academic Publishers, 2004, pp. 53-68.
14. Gorodnitsky I, Belouchrani A., " Joint cumulant and correlation based signal separation with application of EEG data analysis" , in Proc. 3-rd Int. Conf. on Independent Component Analysis and Signal Separation, San Diego, California, Dec. 9-13, 2001, pp.475-480.
15. A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.
16. T.-W. Lee, M.S. Lewicki, M. Girolami, T.J. Sejnowski, "Blind sourse separation of more sources than mixtures using overcomplete representaitons", *IEEE Signal Process. Lett.*, Vol. 6, no. 4, pp. 87–90, 1999.
17. D. D. Lee and H. S. Seung "Learning the parts of objects by non-negative Matrix Factorization", *Nature*, Vol. 40, pp. 788–791, 1999.
18. Iriarte J, Urrestarazu E, Valencia M, Alegre M, Malanda A, Viteri C, Artieda J., "Independent component analysis as a tool to eliminate artifacts in EEG: a quantitative study", J. Clin. Neurophysiol. 2003 Jul-Aug 20:4, pp. 249-57.
19. Jung TP, Makeig S, Westerfield M, Townsend J, Courchesne E, Sejnowski TJ, "Analysis and visualization of single-trial event-related potentials", Hum. Brain Mapp. 2001 Nov, 14:3, pp. 166-85.
20. Krieger S, Timmer J, Lis S, Olbrich HM., "Some considerations on estimating event-related brain signals", *J. Neural Transm. Gen. Sect.* 1995, 99:103, pp 29.
21. McKewon, M., Jung, T., Makeig, S., Brown, G., Kindermann, S., Bell, A., Sejnowski, T., "Analysis of fMRI data by blind separation into independent spatial components", *Human Brain Mapping* 6, 1998, pp. 160–188.

22. M. McKeown, L. Hansen and T. Sejnowski, "Independent component analysis of functional MRI: what is signal and what is noise? *Current Opinion in Neurobiology* 2003, 13, pp. 620-629.
23. Naatanen R, Picton T., "The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure", *Psychophysiology*, 1987, Jul. 24:4, pp. 375-425.
24. Potts GF, Dien J, Hartry-Speiser AL, McDougal LM, Tucker DM., "Dense sensor array topography of the event-related potential to task-relevant auditory stimuli", *Electroencephalogr. Clin. Neurophysiol.*, 1998, May 106:5, pp. 444-56.
25. F.J. Theis, E.W. Lang, and C.G. Puntonet, A geometric algorithm for overcomplete linear ICA. *Neurocomputing*, in print, 2003.
26. K. Waheed, F. Salem, "Algebraic Overcomplete Independent Component Analysis", in *Proc. Int. Conf. ICA2003*, Nara, Japan, pp. 1077–1082.
27. A. Wismüller, O. Lange, D. Dersch, G. Leinsinger, K. Hahn, B. Pütz and D. Auer, "Cluster Analysis of Biomedical Image Time–Series", *International Journal on Computer Vision*, Vol. 46, 2, 2002, pp.102–128.
28. R. Woods and S. Cherry and J. Mazziotta, "Rapid automated algorithm for aligning and reslicing PET images", *Journal of Computer Assisted Tomography*, Vol. 16, 8, 1992, pp. 620–633.
29. M. Zibulevsky, and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary", *Neural Comput.*, Vol. 13, no. 4, pp. 863–882, 2001.