Journal of VLSI Signal Processing 48, 161–171, 2007

DOI: 10.1007/s11265-006-0028-3

Overcomplete Independent Component Analysis via Linearly Constrained Minimum Variance Spatial Filtering

MORITZ GROSSE-WENTRUP AND MARTIN BUSS
Institute of Automatic Control Engineering, Technische Universität München, 80290, München, Germany

Received: 1 March 2006; Revised: 17 August 2006; Accepted: 18 October 2006

Abstract. Independent Component Analysis (ICA) designed for complete bases is used in a variety of applications with great success, despite the often questionable assumption of having N sensors and M sources with $N \geq M$. In this article, we assume a source model with more sources than sensors ($M>N$), only $L<N$ of which are assumed to have a non-Gaussian distribution. We argue that this is a realistic source model for a variety of applications, and prove that for ICA algorithms designed for complete bases (i.e., algorithms assuming $N=M$) based on mutual information the mixture coefficients of the L non-Gaussian sources can be reconstructed in spite of the overcomplete mixture model. Further, it is shown that the reconstructed temporal activity of non-Gaussian sources is arbitrarily mixed with Gaussian sources. To obtain estimates of the temporal activity of the non-Gaussian sources, we use the correctly reconstructed mixture coefficients in conjunction with linearly constrained minimum variance spatial filtering. This results in estimates of the non-Gaussian sources minimizing the variance of the interference of other sources. The approach is applied to the denoising of Event Related Fields recorded by MEG, and it is shown that it performs superiorly to ordinary ICA.

Keywords: independent component analysis, blind source separation, overcomplete, underdetermined, EEG, MEG, denoising, event related fields, event related potentials, beamforming

1. Introduction

Independent Component Analysis (ICA) has become a widely used tool in a variety of signal processing applications. By linearly decomposing measured data into maximally independent components (ICs), the analysis of a single source can be decoupled from all other sources [5]. This is utilized in a variety of applications, e.g., temporal analysis [14] and source localization [20] of MEG/EEG data, or the analysis of fMRI data [15] to mention just a few.

For ICA to be applicable to a given data set, four basic requirements have to be fulfilled: (1) a linear mixture model; (2) mutual statistical independence of the original sources; (3) stationarity of the source distributions; (4) at least as many sensors as sources (ICA designed for complete bases). Frequently, it is also stated that only one of the sources may be Gaussian distributed. This is not a necessary requirement if only sources with non-Gaussian distributions are of interest, because these can be separated in spite of the presence of multiple Gaussian sources [10].

The requirement of having at least as many sensors as sources has initiated development of algorithms for overcomplete ICA, i.e., algorithms that can deal with more sources than sensors. For overcomplete ICA, it has been shown that the mixture coefficients of the sources are identifiable under certain conditions, while the sources are not [8]. This necessitates the introduction of a regularization parameter to obtain unique estimates of the sources, e.g., by requiring the sources to be sparse [13].

Interestingly, one of the most widely used ICA algorithms designed for complete bases, the extended Infomax-algorithm [12], is nevertheless applied with seemingly great success to data sets for which more sources than sensors are to be expected, such as in MEG/EEG data analysis.

In MEG/EEG, the continuous current distribution inside the brain is recorded by a finite number of sensors. This corresponds to a mapping from an infinite to a finite dimensional space. This mapping could only have a unique inverse or be underdetermined, as it is usually assumed in studies applying ICA to MEG/EEG (c.f. [11]), if the continuous current distribution inside the brain could be partitioned into N or less sets with absolutely identical temporal activity, where N is the number of sensors. From a physiological point of view this is highly unlikely, and to the best of our knowledge there is no empirical evidence supporting this assumption. We rather maintain that this assumption has only been adopted as a working hypothesis to justify applying ICA designed for complete bases to MEG/EEG data.

In this article, we investigate the performance of ICA algorithms designed for complete bases based on mutual information, such as the extended Infomax-algorithm, for cases where more sources than sensors are present. More specifically, we assume an overcomplete mixture model, but assume fewer sources with non-Gaussian distributions than sensors. This source model is based on (a) the fact that only sources with a non-Gaussian distribution can be consistently reconstructed with algorithms based on mutual information, and (b) the observation that for a variety of applications typically only a few ICs can be consistently reconstructed, i.e., independent of initial conditions of the algorithm (see [10] and references therein). For such data sets, which include MEG/EEG, it can be concluded that fewer non-Gaussian sources than sensors are present.

We prove that for this source model, ICA algorithms designed for complete bases based on mutual information can correctly identify the mixture coefficients of the non-Gaussian sources, but arbitrarily mix Gaussian sources into the reconstructed temporal activity of non-Gaussian sources. To obtain optimal estimates of the temporal source activity of the non-Gaussian sources, we formulate an optimization problem based on linearly constrained minimum variance spatial filtering (LCMV) [19]: For each non-Gaussian distributed source, we find a linear transformation that minimizes the overall variance, under the constraint of the product of the linear transformation with the correctly identified mixture coefficients of the respective source being unity. This leads to source estimates of the non-Gaussian sources that minimize the variance of the interference of all other sources. We apply this approach to the denoising of Auditory Event Related Fields (AEFs) recorded by MEG, and show that it performs superiorly to ordinary ICA.

It should be pointed out that other authors previously suggested combining Blind Source Separation (BSS) with spatial filtering (a general overview of spatial filtering in the context of MEG/EEG analysis can be found in [9]). In fact, one of the first studies on BSS proposed to use blind identification in the context of beamforming to address inaccuracies in the physical model of array manifolds [4]. More recent studies include [17], addressing ambiguities in convolutive source separation by geometric beamforming, and [18], combining ICA and beamforming to obtain better convergence properties. All of these studies, however, are restricted to complete mixture models. To the best of our knowledge, this study is the first to address ICA and beamforming in the context of overcomplete mixture models.

The rest of the article is organized as follows. In Section 2, we first introduce the overcomplete source model and ICA based on mutual information, followed by the proof that the mixture coefficients of non-Gaussian sources are identifiable, while the temporal activity of non-Gaussian sources is arbitrarily mixed with Gaussian sources. We then show how optimal estimates of the non-Gaussian sources can be obtained by minimizing the interference of all other sources. In the Section 3, we apply our approach to AEFs recorded by MEG, and compare the results with ordinary ICA. We conclude with a discussion of the implications of the results.

2. Materials and Methods

2.1. Overcomplete Source Model

The overcomplete source model is given by

$$x = As, \tag{1}$$

with the random variables $x \in \mathbb{R}^N$, $s \in \mathbb{R}^M$ with $M > N$. The matrix $A \in \mathbb{R}^{N \times M}$ is assumed to have full row-rank, and the sources s_i, $i = 1 \ldots M$ are assumed to be mutually statistically independent, i.e.,

$$p(s) = \prod_{i=1}^{M} p(s_i). \tag{2}$$

Furthermore, it is assumed that $s_i \sim N(0,1)$, $i = L+1 \ldots M$, i.e., only the first L sources are assumed to have a non-Gaussian distribution. Without loss of generality, all sources are assumed to have zero mean and unit variance. Additive measurement noise is included in this model as a special case, i.e., as sources only projecting to one sensor. Without loss of generality, we furthermore assume that the measurements x have been sphered, i.e., that the data covariance matrix is given by

$$R_x = \langle x, x \rangle = A \langle s, s \rangle A^{\mathrm{T}} = AA^{\mathrm{T}} = I_{N \times N}. \tag{3}$$

The N rows of the mixing matrix A are hence mutually orthogonal.

2.2. Complete ICA for Overcomplete Source Models

To reconstruct the temporal source activity s from the measurements in the case of ICA designed for complete bases, we search for a matrix W such that

$$y = Wx = WAs = s \tag{4}$$

with $W \in \mathbb{R}^{N \times N}$. Obviously this problem is ill-posed for the overcomplete mixture model, since we are trying to find a one-to-one mapping from a M-dimensional to a N-dimensional vectorspace with $M > N$. It is thus evident, that the original M sources cannot be separated from the N measurements.

We will now investigate the form of unmixing matrices W as well as the reconstructed source vector y returned by ICA algorithms designed for complete bases based on mutual information if applied to this ill-posed problem. For this purpose, first note that since x is sphered, only orthogonal unmixing matrices have to be considered [3]. The estimated sources are thus obtained by

$$y = Wx \tag{5}$$

with $W \in R^{N \times N}$ orthogonal. For algorithms based on mutual information, the matrix W is found by minimizing the mutual information between the elements of y:

$$\min_{W} \left\{ \sum_{i=1}^{N} H(y_i) - H(y) \right\} \tag{6}$$

with $H(y_i) = -\int_{-\infty}^{\infty} p_{y_i}(u) \log(p_{y_i}(u)) du$ the differential entropy. Since W is an invertible transformation, Eq. (6) can be rewritten as

$$\min_{W} \left\{ \sum_{i=1}^{N} H(y_i) - \log(|W|) - H(x) \right\} \tag{7}$$

which reduces to

$$\min_{W} \left\{ \sum_{i=1}^{N} H(y_i) \right\} \tag{8}$$

since W is orthogonal and $H(x)$ is independent of W.

The following derivations extend the results of [1] to the overcomplete case. Define $F(W) := \sum_{i=1}^{N} H(y_i)$. The gradient of $F(W)$ under the orthogonality constraint then becomes [7]

$$\nabla_{\mathrm{ortho}} F(W) = \nabla F(W) - W \nabla F(W)^T W. \tag{9}$$

Since $WW^T = I_{N \times N}$, solutions of Eq. (9) are given by

$$\nabla F(W) W^T = W \nabla F(W)^T. \tag{10}$$

Note, however, that Eq. (10) is only a necessary condition for a minimum of Eq. (8). Denoting $h(w_i) := H(y_i)$ with w_i the ith row of W, Eq. (10) becomes

$$\nabla h(w_k) \cdot w_l^T = \nabla h(w_l) \cdot w_k^T \quad (11)$$

for $k, l = 1 \ldots N,\ k \neq l$. With

$$\frac{\partial h(y_i)}{\partial w_{i,j}} = - \int_{-\infty}^{\infty} \left(\log p_{y_i}(u) + 1\right) \frac{\partial p_{y_i}(u)}{\partial w_{i,j}} du, \quad (12)$$

Eq. (11) results in

$$\begin{aligned} &\int_{-\infty}^{\infty} \left(\log p_{y_k}(u) + 1\right) \\ &\quad \times \left[\frac{\partial p_{y_k}(u)}{\partial w_{k,1}} w_{l,1} + \ldots + \frac{\partial p_{y_k}(u)}{\partial w_{k,N}} w_{l,N}\right] du \\ &= \int_{-\infty}^{\infty} \left(\log p_{y_l}(u) + 1\right) \\ &\quad \times \left[\frac{\partial p_{y_l}(u)}{\partial w_{l,1}} w_{k,1} + \ldots + \frac{\partial p_{y_l}(u)}{\partial w_{l,N}} w_{k,N}\right] du \end{aligned} \quad (13)$$

for $k, l = 1 \ldots N,\ k \neq l$. A sufficient condition for Eq. (13) to hold is

$$\frac{\partial p_{y_k}(u)}{\partial w_{k,1}} w_{l,1} + \ldots + \frac{\partial p_{y_k}(u)}{\partial w_{k,N}} w_{l,N} = 0 \quad (14)$$

for $k, l = 1 \ldots N,\ k \neq l$.

To simplify the analysis of Eq. (14), note that

$$y = WAs. \quad (15)$$

Defining $C := WA \in R^{N \times M}$, we can write the elements of y as

$$y_i = c_{i,1} s_1 + \ldots + c_{i,M} s_M, \quad (16)$$

with $c_{i,j}$ denoting the element of C in the i^{th} row and the j^{th} column. The probability distribution of y_i is then given by

$$p_{y_i}(u) = \frac{1}{c_{i,1}} p_{s_1}\left(\frac{u}{c_{i,1}}\right) * \ldots * \frac{1}{c_{i,M}} p_{s_M}\left(\frac{u}{c_{M,1}}\right). \quad (17)$$

The analysis of Eq. (14) is further simplified in the frequency domain. With $\varphi_{y_i}(\omega)$ the characteristic function of p_{y_i}, Eq. (17) becomes

$$\varphi_{y_i}(\omega) = \varphi_{s_i}(c_{i,1}\omega) \times \ldots \times \varphi_{s_M}(c_{M,1}\omega). \quad (18)$$

Substituting Eq. (18) in Eq. (14) and dividing by $\varphi_{s_i}(c_{i,1}\omega) \times \ldots \times \varphi_{s_M}(c_{M,1}\omega)$ results in

$$\frac{\omega \varphi'_{s_1}(c_{k,1}\omega) c_{l,1}}{\varphi_{s_1}(c_{k,1}\omega)} + \ldots + \frac{\omega \varphi'_{s_M}(c_{k,M}\omega) c_{l,M}}{\varphi_{s_M}(c_{k,M}\omega)} = 0 \quad (19)$$

for $k, l = 1 \ldots N,\ k \neq l$. Now only if s_i has a Gaussian distribution it holds that

$$\varphi'_{s_i}(\alpha\omega) = -\alpha\omega\varphi_{s_i}(\alpha\omega). \quad (20)$$

Since the sources $s_i,\ i = L+1 \ldots M$ are assumed to be Gaussian, Eq. (19) simplifies to

$$\begin{aligned} &\frac{\omega \varphi'_{s_1}(c_{k,1}\omega) c_{l,1}}{\varphi_{s_1}(c_{k,1}\omega)} + \ldots + \frac{\omega \varphi'_{s_L}(c_{k,L}\omega) c_{l,L}}{\varphi_{s_N}(c_{k,L}\omega)} \\ &- \omega^2 (c_{k,L+1} c_{l,L+1} + \ldots + c_{k,M} c_{l,M}) = 0 \end{aligned} \quad (21)$$

for $k, l = 1 \ldots N,\ k \neq l$. Note that $\varphi'_{s_i}(c_{i,j}\omega)|_{c_{i,j}=0} = 0$ because all sources have zero mean. Considering the first two terms of Eq. (21), the first term is zero if and only if in the first column of C for every pair of elements only one of them is non-zero. This in turn implies that only one element of each column may be non-zero. The same holds for every up to and including the L^{th} column. Considering the last term of Eq. (21), this term is zero if the rows of C, starting with the $(L+1)^{th}$ element, are mutually orthogonal. Now note that the rows of C are mutually orthogonal, since $C = WA$ and W is orthogonal and the rows of A are mutually orthogonal. Hence, the rows of C may only have on non-zero entry in the first L columns. In summary, it is sufficient for Eq. (21) to hold that C is of the form

$$C = (P_{N \times L} | Q_{N \times M-L}) \quad (22)$$

with $P \in \mathbb{R}^{N \times L}$ any permutation matrix (only one non-zero entry in each column and row) and $Q \in \mathbb{R}^{N \times M-L}$ any matrix with mutually orthogonal rows.

We will now investigate what this implies for the structure of the unmixing matrix W and the estimated mixing matrix $\widetilde{A} = W^{-1}$. Without loss of generality, we assume that no permutation and scaling of the sources takes place, i.e.,

$$C = WA = (I_{N\times L}|Q_{N\times M-L}) \tag{23}$$

with $I_{N\times L}$ having unit entries for the first L diagonal elements and zero entries otherwise. With w_i the i^{th} row of W and a_j the j^{th} column of A, this implies that

$$w_i a_j = 0 \tag{24}$$

for $i = 1\ldots N,\ j = 1\ldots L$ and $i \neq j$, and

$$w_i a_j \neq 0 \tag{25}$$

otherwise. This means that the i^{th} row of W is orthogonal to all column vectors of A representing the non-Gaussian sources except the i^{th} column of A, but not orthogonal to the column vectors of A representing the Gaussian sources. For $i \leq L$, the i^{th} row of W thus lies in a subspace that is orthogonal to the subspace spanned by all non-Gaussian sources except the i^{th} non-Gaussian source.

Now consider the estimated mixing matrix $\widetilde{A} = W^{-1}$. With

$$W\widetilde{A} = I_{N\times N}, \tag{26}$$

by construction, we see that the j^{th} column of $\widetilde{A}$ is orthogonal to all except the j^{th} row of W. Consequently, the j^{th} column of $\widetilde{A}$ lies in a subspace that is orthogonal to all except the j^{th} row of W. Now note that as discussed above for $j = 1\ldots L$, the rows w_i of W with $i = 1\ldots N, i \neq j$ span the subspace of all sources except the j^{th} non-Gaussian source. Consequently, the j^{th} column of $\widetilde{A}$ is linearly dependent on the j^{th} column of A for $j = 1\ldots L$, and thus correctly identifies the mixing coefficients of the original non-Gaussian source up to multiplication by a constant.

We can summarize the results of this section as follows. For the overcomplete mixing model with N sensors, M sources out of which only L have a non-Gaussian distribution, and $M > N > L$, any unmixing matrix W that fulfills $WA = C$ with C in the form of Eq. (22) is a stationary point of the minimization problem (8). The reconstructed sources are then given by $y = Wx = WAs = Cs$. Thus, algorithms for complete ICA based on mutual information are capable of separating the non-Gaussian sources, but Gaussian sources can be arbitrarily mixed into the reconstructed non-Gaussian sources. However, even if the non-Gaussian sources are not separated from the Gaussian sources, the estimated mixing matrix $\widetilde{A}$ correctly reconstructs the columns of the original mixing matrix corresponding to the non-Gaussian sources up to permutation and scaling. In the next section, we will show how the correctly identified columns of the mixing matrix corresponding to the non-Gaussian sources can be used to estimate the non-Gaussian sources in an optimal manner.

2.3. *Source Estimation by Linearly Constrained Minimum Variance Filtering*

As we have seen in the previous section, for the assumed source model ICA algorithms based on mutual information, such as the extended Infomax-algorithm, arbitrarily mix Gaussian sources into the reconstructed temporal activity of non-Gaussian sources. The correct identification of the mixture coefficients of the sources with non-Gaussian distributions, however, enables an optimal estimation of the temporal activity of these sources. In this section, we show how the temporal activity of the non-Gaussian sources can be estimated by minimizing the variance of the interference of all other sources.

Consider again the overcomplete mixture model given in Eq. (1), but this time assume that the columns $a_i,\ i = 1\ldots L$, corresponding to the mixture coefficients of the non-Gaussian sources, to be known. To estimate the temporal activity $\widetilde{s}_i$ of the i^{th} source, we wish to find a linear transformation v_i that passes all activity originating from the i^{th} source, while attenuating all other sources:

$$\widetilde{s}_i = v_i^T x. \tag{27}$$

If we choose to minimize the variance of all except the i^{th} source, v_i can be found by solving the following optimization problem:

$$\min_{v_i}\{\widetilde{s}_i^{\,2}\} \text{ s.t. } v_i^{\mathrm{T}} a_i = 1, \tag{28}$$

which can be rewritten as

$$\min_{v_i}\{v_i^T R_x v_i\} \text{ s.t. } v_i^T a_i = 1, \tag{29}$$

with R_x the (estimated) data covariance matrix. The solution to this optimization problem is given in [19]:

$$v_i = (a_i^T R_x^{-1} a_i)^{-1} a_i^T R_x^{-1}. \tag{30}$$

If a_i corresponds to the mixing coefficients of a non-Gaussian source, the resulting $\tilde{s}_i$ is an estimate of the original non-Gaussian source with minimized variance of the interference of all other sources. Note, however, that this also implies that the estimated source activity is not statistically independent of all other non-Gaussian sources anymore: statistical independence is traded for minimization of the variance of interference of all other sources. If a_i corresponds to a source with Gaussian distribution, the estimated $\tilde{s}_i$ is meaningless, in the sense that the reconstructed activity corresponds to an arbitrary mixture of original sources with Gaussian distribution. Whether a_i corresponds to a source with non-Gaussian or Gaussian distribution cannot be determined from the mixture coefficients, but has to be deduced from the estimated temporal source activity $\tilde{s}_i$. Also note that in the derivation of Eq. (30) the original sources are assumed to be uncorrelated, which is fulfilled for the source model considered here due to the mutual statistical independence assumption.

3. Results

To evaluate the efficacy of the approach proposed in Section 2, we apply it to the denoising of MEG data by ICA in this section, and compare its performance with ordinary ICA. Data denoising by ICA is based on the assumption that only a small number of ICs reconstructed from a given data set are relevant for the considered experimental setup, i.e., belong to the signal subspace, while all other ICs constitute noise. Only the ICs belonging to the signal subspace are then reprojected onto the observation space, resulting in a rank-reduced signal with improved signal-to-noise ratio (SNR). It should be noted that the identification of ICs relevant for a given experimental setup is not trivial, and hence mostly done manually. In the context of the source model considered here, we assume that only the L non-Gaussian sources belong to the signal subspace. We hence consider the deviation from Gaussianity of the reconstructed sources as a criterion for the identification of relevant ICs.

As MEG data we chose Event Related Fields (ERFs). ERFs typically have a very low SNR, and are difficult to detect in single trial data. For this reason numerous trials are recorded, and the ERF is estimated by taking the ensemble average of all trials. Based on the assumption that only the ERF component of the MEG is invariant in every trial, this results in an unbiased estimator of the ERF (termed the *grand average* ERF). In complex experimental setups, or if subjects with a short attention span such as small children are under investigation, the recording of numerous trials is not feasible. The goal of ERF denoising by ICA is then to reconstruct the grand average ERF from only a small number of trials. This application is well suited for evaluating the approach presented in Section 2, because a data set can be used for which the grand average ERF actually is available. This allows an objective evaluation of the obtained denoising results. Furthermore, as we have argued in the introduction, the overcomplete source model (1) is a realistic assumption for MEG data sets.

The test data set consists of Auditory Evoked Fields (AEFs), recorded during an auditory oddball task at the Biomagnetic Imaging Laboratory of the University of California, San Francisco. Auditory stimuli were applied to the left ear, while MEG was recorded at a sampling rate of 4 kHz with $N = 132$ sensors covering the right hemisphere. A total of 250 trials were recorded, with each trial lasting from -275 to 275 ms and the stimulus being applied at 0 ms (see [16] for a detailed description of the recording procedure). Out of the total number of 250 trials, ten trials were chosen randomly for estimation of the raw average ERF. The grand average y^* was computed by taking the average time course of all 250 trials, and filtering the resulting average sequentially with a low- and high-pass filter with cut-off frequencies 2 Hz and 16 Hz , respectively, (for all temporal filtering procedures in this article a third order Butterworth filter was used). The resulting temporal activity at all channels is shown in Fig. 1a. The same temporal filtering procedure was applied to the average of the randomly chosen ten trials, resulting in the temporal activity y^{raw} shown in Fig. 1b. Note that only the post-stimulus period is shown in both figures. For a quantitative comparison of the data sets, the SNR was defined as

$$\mathrm{SNR}(\widehat{y}) := 10\log_{10}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{t=1}^{T}y_i^*[t]^2}{\sum_{t=1}^{T}\left(y_i^*[t]-\widehat{y}_i[t]\right)^2}\right)(dB), \tag{31}$$

with samples $t = 1\ldots T$ corresponding to the post-stimulus period of the data. Each data sets was first normalized to the maximum value of all channels before computing the SNR. This resulted in a SNR of -0.09 dB for the data set y^{raw}.

To evaluate the denoising capabilities of ICA, the extended Infomax-algorithm as implemented in EEGLab [6] was applied to the concatenated ten trials that were randomly chosen as test data (from here on referred to as the data vector x), resulting in estimated source topographies $\widehat{a}_i$ and temporal source estimates $\widehat{s}_i$ with $i = 1, \ldots, N$. Four different evaluation schemes were then investigated:

1. *Ordinary ICA* The reconstructed sources $\widehat{s}_i$ are sorted in descending order according to the variance of the original data explained by each source. Only the first L sources with the highest explained variance are reprojected onto the observation space,

$$\widehat{x}^1 = \sum_{i=1}^{L}\widehat{a}_i\widehat{s}_i. \tag{32}$$

2. *ICA with LCMV spatial filtering* The temporal source activity of each source is estimated using the LCMV spatial filtering approach (27), (28), (29), (30), and the resulting source estimates are again sorted in descending order according to the amount of variance of the original data explained by each source. The first L sources explaining the highest amount of variance are reprojected onto the observation space, resulting in

$$\widehat{x}^2 = \sum_{i=1}^{L}\widehat{a}_i(\widehat{a}_i^{\mathrm{T}}R_x^{-1}\widehat{a}_i)^{-1}\widehat{a}_i^{\mathrm{T}}R_x^{-1}x. \tag{33}$$

Note that in this and the fourth evaluation scheme diagonal loading is used to obtain numerically stable estimates of the inverse of the covariance matrix R_x.

3. *Ordinary ICA with identification of relevant non-Gaussian sources* The sources are reconstructed with ordinary ICA, but not sorted in descending order according to the amount of variance explained by each IC. Instead, the

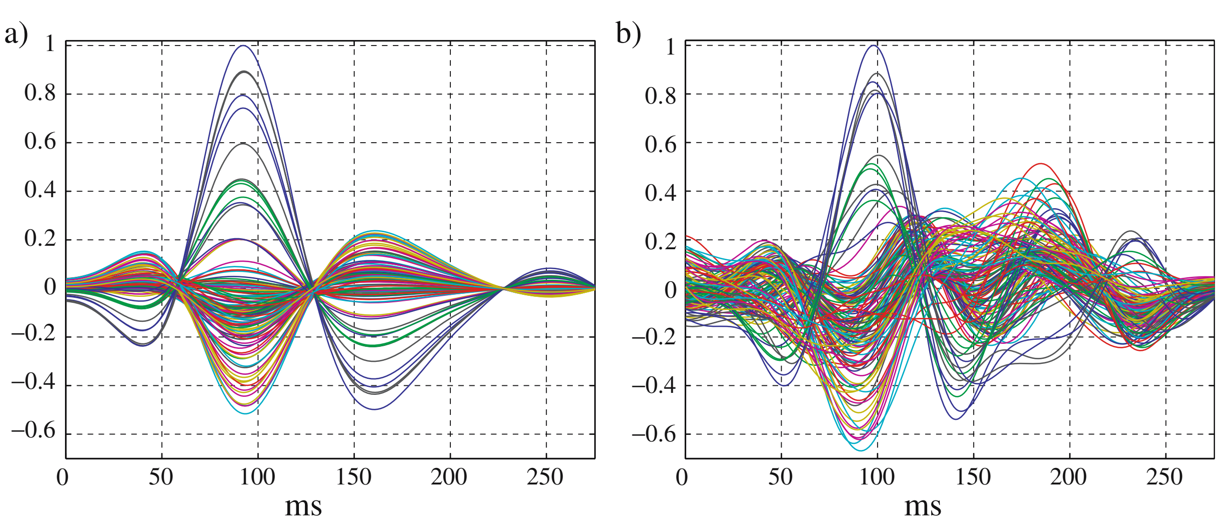

Figure 1. Grand average ERF y^* (**a**) and ERF average of ten randomly chosen trials y^{raw} (**b**).

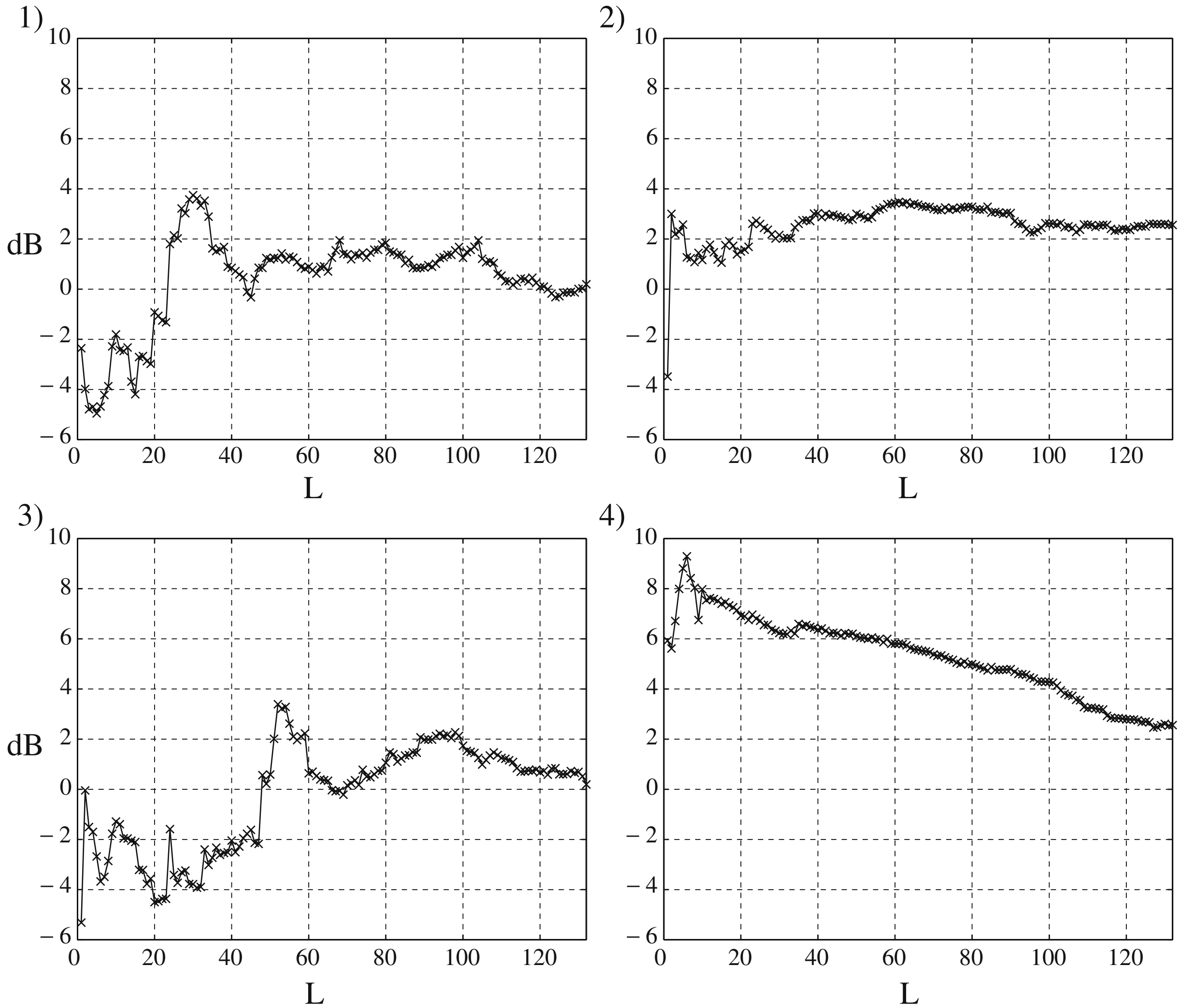

Figure 2. SNR of the evaluation schemes 1–4.

deviation from Gaussianity of each source $\widehat{s}_i$ is estimated in multiple stages. First, the average temporal activity of each source across the ten trials is computed. Then, the probability distribution function (pdf) of each averaged source is estimated for the post-stimulus period using a non-parametric kernel approach (c.f. [2]). A Gaussian kernel is used, which is optimal for Gaussian distributions. Then, the Kullback-Leibler distance of the estimated pdf to a Gaussian distribution with equal variance is calculated by numerical integration. Finally, the sources are sorted from highest to lowest Kullback-Leibler distance, i.e., from least to most Gaussian. The data set $\widehat{x}^3$ is then calculated in the same way as in Eq. (32), but by reprojecting the L most non-Gaussian sources.

4. *ICA with LCMV spatial filtering and identification of relevant non-Gaussian sources* The temporal source activity of each source is again estimated

Table 1. Maximum SNR for each of the four denoising schemes.

Evaluation scheme	1	2	3	4
Maximum SNR	3.75 dB	3.48 dB	3.32 dB	9.29 dB
L_{max}	30	61	51	6

using Eqs. (27), (28), (29), (30). The estimated sources are sorted in descending order according to their deviation from Gaussianity as for evaluation scheme three. The data set $\widehat{x}^4$ is then calculated in the same way as in Eq. (33), but by reprojecting the L most non-Gaussian sources.

The denoised data sets $\widehat{y}^j,\ j = 1, \ldots, 4$ are calculated from the data sets $\widehat{x}^j$ by taking the average across the ten trials of x^j, applying the same temporal filtering procedure as for the grand average data set, and normalizing to the maximum value across all channels of each $\widehat{y}^j$. Note that determining the parameter L, corresponding to the dimension of the signal subspace, is a non-trivial issue related to model identification. This is beyond the scope of this article. The resulting SNRs for all four schemes applied to the ten randomly chosen trials are shown in Fig. 2 in dependence on the choice of L. The maximum SNR achieved for each evaluation scheme is summarized in Table 1, with Fig. 3 showing the corresponding time series.

As can be seen from Table 1, the best SNR of 9.29 dB is achieved for ICA with LCMV spatial filtering and sorting of the estimated sources by their deviation from Gaussianity. The SNRs for the other three evaluation schemes are roughly equal at about 3.5 dB. Note that the best SNR for evaluation

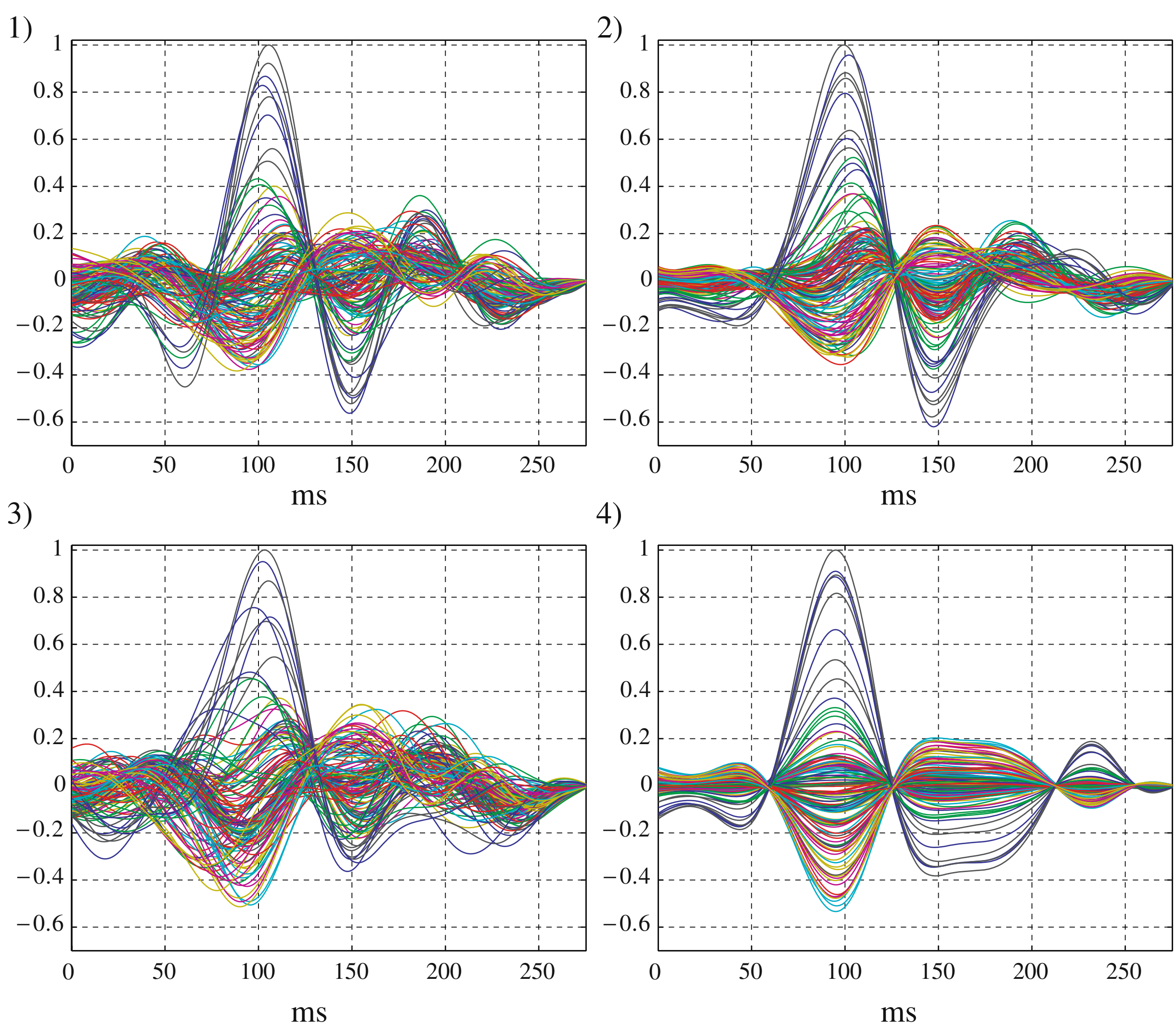

Figure 3. Denoised ERFs with optimal L for evaluation schemes 1–4.

scheme 4 is obtained for $L = 6$, while the optimal SNR for the other evaluation schemes is obtained for much higher dimensions of the signal subspace (Fig. 2). As it can be expected from the SNRs, the temporal activity at the recording channels for the optimum SNR of each evaluation scheme differs significantly (Fig. 3). While the 4^{th} evaluation scheme correctly reconstructs all major peaks of the grand average ERF (see Fig. 1), for the other three evaluation schemes only the major peak around 100 ms is clearly discernible.

4. Discussion

Summarizing the experimental results, it was shown that ICA with estimation of the temporal source activity by linearly constrained minimum variance spatial filtering and identification of relevant sources by deviation from Gaussianity is superior to ordinary ICA. This is in agreement with the theoretical results presented in the Section 2, proving that ICA designed for complete bases based on mutual information can correctly identify the mixture coefficients of the non-Gaussian sources, but arbitrarily mixes reconstructed non-Gaussian with Gaussian sources. The experimental results furthermore support the argument that the overcomplete source model with less non-Gaussian sources than sensors is a realistic assumption for MEG data. We thus conclude that great care should be taken in the interpretation of temporal activity of sources reconstructed by ordinary ICA if applied to data sets for which the overcomplete source model presented here is a realistic assumption, such as MEG or EEG data. In these cases, the reconstructed temporal activity of a source can be arbitrarily mixed with Gaussian sources. The experimental results show that for non-Gaussian distributed sources, estimation of the temporal source activity by linearly constrained minimum variance spatial filtering results in improved estimates of the original temporal source activity.

Acknowledgement

The authors would like to thank the anonymous reviewers for the helpful comments.

References

1. R. Boscolo, H. Pan, and V. P. Roychowdhury, "Beyond Comon's Identifiability Theorem for Independent Component Analysis," *Artificial Neural Networks—ICANN 2002 Lecture Notes in Computer Science*, vol. 2415, 2002, pp. 1119–1124.
2. A. W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis*, Oxford University Press, 1997.
3. J. F. Cardoso, "Blind Signal Separation: Statistical Principles," *Proc. IEEE*, vol. 86, 1998, pp. 2009–2025.
4. J. F. Cardoso and A. Souloumiac, "Blind Beamforming for Non–Gaussian Signals," *IEE Proc. F*, vol. 140, 1993, pp. 362–370.
5. P. Comon," Independent Component Analysis, A New Concept?," *Signal Process.*, vol. 36, 1994, pp. 287–314.
6. A. Delorme and S. Makeig, "EEGLAB: An Open Source Toolbox for Analysis of Single-trial EEG Dynamics," *J. Neurosci. Methods*, vol. 134, 2004, pp. 9–21.
7. A. Edelman, T. A. Arias, and S. T. Smith, "The Geometry of Algorithms with Orthogonality Constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, 1998, pp. 303–353.
8. J. Eriksson and V. Koivunen, "Identifiability and Separability of Linear ICA Models Revisited," in *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2003)*, 2003, pp. 23–27.
9. J. Gross and A. A. Ioannides, "Linear Transformations of Data Space in MEG," *Phys. Med. Biol.*, vol. 44, 1999, pp. 2081–2097.
10. M. Grosse-Wentrup and M. Buss, "Subspace Identification through Blind Source Separation," *IEEE Signal Process. Lett.*, vol. 13, no. 2, 2006, pp. 100–103.
11. C. J. James and O. J. Gibson, "Temporally Constrained ICA: An Application to Artifact Rejection in Electromagnetic Brain Signal Analysis," *IEEE Trans. Biomed. Eng.*, vol. 50, 2003, pp. 1108–1116.
12. T. W. Lee, M. Girolami, and T. J. Sejnowski, "Independent Component Analysis Using an Extended Infomax Algorithm for Mixed Subgaussian and Supergaussian Sources," *Neural Comput.*, vol. 11, 1999, pp. 417–441.
13. T. W. Lee, M. S. Lewicki, M. Girolami, and T. J. Sejnowski, "Blind Source Separation of More Sources Than Mixtures Using Overcomplete Representations," *IEEE Signal Process. Lett.*, vol. 6, 1999, pp. 87–90.
14. S. Makeig, M. Westerfield, T. P. Jung, S. Enghoff, J. Townsend, E. Courchesne, and T. J. Sejnowski, "Dynamic Brain Sources of Visual Evoked Responses," *Science*, vol. 295, 2002, pp. 690–694.
15. M. J. McKeown, S. Makeig, G. G. Brown, T. P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski, "Analysis of fMRI Data by Blind Separation into Independent Spatial Components," *Hum. Brain Mapp.*, vol. 6, 1998, pp. 160–188.
16. S. S. Nagarajan, H. T. Attias, and K. E. Hild II, and K. Sekihara, "A Graphical Model for Estimating Stimulus-evoked Brain Responses from Magnetoencephalography Data with Large Background Brain Activity," *NeuroImage*, vol. 30, 2006, pp. 400–416.
17. L. C. Parra and C. V. Alvino, "Geometric Source Separation: Merging Convolutive Source Separation with Geometric Beamforming," *IEEE Trans. Speech Audio Process.*, vol. 10, 2002, pp. 352–362.

18. H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind Source Separation Based on a Fast-convergence Algorithm Combining ICA and Beamforming," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, 2006, pp. 666—678.
19. B. D. van Veen, W. van Drongelen, M. Yuchtman, and A. Suzuki, "Localization of Brain Electrical Activity via Linearly Constrained Minimum Variance Spatial Filtering," *IEEE Trans. Biomed. Eng*., vol. 44, 1997, pp. 867—880.
20. L. Zhukov, D. Weinstein, and C. Johnson, "Independent Component Analysis for EEG Source Localization—An Algorithm that Reduces the Complexity of Localizing Multiple Neural Sources," *IEEE Eng. Med. Biol. Mag*., vol. 19, 2000, pp. 87–96.

Dipl.-Ing. Moritz Grosse-Wentrup obtained his diploma in Electrical Engineering and Information Technology from the Technical University Munich, Germany, in 2004. He is currently pursuing a Ph.D. in electrical engineering and information technology with funding from the German National Academic Foundation at the Institute of Automatic Control Engineering, Technische Universitaet Muenchen. His research interests include: Signal processing, independent component analysis, brain-machine interfaces.

Martin Buss was born in Germany in 1965. He received the diploma engineer degree in electrical engineering in 1990 from the Technical University Darmstadt, Germany, the Doctor of Engineering degree in electrical engineering from the University of Tokyo, Japan, in 1994, and the habilitation degree from the Department of Electrical Engineering and Information Technology, Technical University Munich, Germany.

In 1988 he was a research student at the Science University of Tokyo, Japan, for one year. As a postdoctoral researcher he stayed with the Department of Systems Engineering, Australian National University, Canberra, Australia, in 1994/5. From 1995-2000 he has been senior research assistant and lecturer at the Institute of Automatic Control Engineering, Department of Electrical Engineering and Information Technology, Technical University Munich, Germany. From 2000-2003 he was full professor, head of the control systems group, and deputy director of the Institute of Energy and Automation Technology, Faculty IV – Electrical Engineering and Computer Science, Technical University Berlin, Germany. Since 2003 he is full professor, chair, and director of the Institute of Automatic Control Engineering and Autonomous Systems, Department of Electrical Engineering and Information Technology, Technical University Munich (TUM), Munich, Germany.

His research interests include control, mechatronics, multi-modal human-system interfaces, optimization, nonlinear, and hybrid discrete-continuous systems.