BLIND SOURCE SEPARATION TECHNIQUES
FOR DECOMPOSING EVENT-RELATED
BRAIN SIGNALS

KLAUS-ROBERT MÜLLER*†
Department of Computer Science, University of Potsdam,
August-Bebel-Strasse 89, 14482 Potsdam, Germany
klaus@first.fhg.de

RICARDO VIGÁRIO‡
Neural Networks Research Centre, HUT, P.O. Box 5400, FIN-02015 HUT, Finland
Ricardo.Vigario@hut.fi

FRANK MEINECKE‡
Department of Physics, University of Potsdam,
Am Neuen Palais 10, 14469 Potsdam, Germany
meinecke@first.fhg.de

ANDREAS ZIEHE
‡Fraunhofer FIRST.IDA, Kekuléstr. 7, 12489 Berlin, Germany
ziehe@first.fhg.de

Received May 15, 2002; Revised November 1, 2002

Recently blind source separation (BSS) methods have been highly successful when applied to biomedical data. This paper reviews the concept of BSS and demonstrates its usefulness in the context of event-related MEG measurements. In a first experiment we apply BSS to artifact identification of raw MEG data and discuss how the quality of the resulting independent component projections can be evaluated. The second part of our study considers averaged data of event-related magnetic fields. Here, it is particularly important to monitor and thus avoid possible overfitting due to limited sample size. A stability assessment of the BSS decomposition allows to solve this task and an additional grouping of the BSS components reveals interesting structure, that could ultimately be used for gaining a better physiological modeling of the data.

Keywords: Blind Source Separation (BSS); Independent Component Analysis (ICA); temporal decorrelation; TDSEP; high order statistics; MEG; evoked responses; reliability; bootstrap.

1. Introduction

Recent advances in brain imaging techniques allow to monitor the active brain noninvasively at high spatial and temporal resolution. Processing this vast amount of data calls for efficient and robust mathematical algorithms to extract meaningful features of interest, while suppressing disturbances due to technical or physiological noise sources. Latent variable models combined with projection techniques are good candidates to achieve this goal. Well-known approaches of this kind are

*Author for correspondence.
e.g. principal component analysis (PCA), projection pursuit (PP) or factor analysis (FA); In this paper we focus on a similar concept known as blind source separation [Jutten & Herault, 1991; Comon, 1994; Hyvärinen et al., 2001].

This method has proven its success in many application areas including e.g. hyperspectral data analysis [Parra et al., 2000], separation of acoustic signals [Lee et al., 1997; Lee et al., 1998; Parra & Spence, 2000; Murata et al., 2001] and the so far most successful application is biomedical data like Electroencephalograms (EEG) [Makeig et al., 1997; Vigario, 1997; Barros et al., 2000], Magnetoencephalograms (MEG) [Vigário, 1999; Vigário et al., 2000; Wübbeler et al., 2000; Ziehe et al., 2001; Sander et al., 2002], Magnetoneurograms (MNG) [Ziehe et al., 2000a], functional magnetic resonance imaging (fMRI) [McKeown et al., 1998; Biswal & Ulmer, 1999; Petersen et al., 2000] or optical (calcium) imaging [Stetter et al., 2000] (see also [Hyvärinen et al., 2001] for further applications). However, as for any unsupervised method, care has to be taken to evaluate the results of such an analysis, especially when studying large complex systems like the human brain.

In this paper we briefly review the BBS concept and discuss its practical use in two biomedical applications: (a) artifact identification in raw MEG measurements and (b) analysis of event-related averaged MEG data.

In Sec. 2 we present a short description of the blind source separation (BSS) problem. The independent component analysis (ICA) theory is then introduced, together with some popular ICA algorithms. Section 3.2 explains a further interesting type of BSS algorithm which uses the temporal structure in the data for unmixing. In Sec. 4 we discuss built-in limits of BSS algorithms (assumptions and statistical overlearning problems) and subsequently compare ICA and PCA briefly. Since BSS is an unsupervised algorithm where one rarely has access to the ground truth, it is mandatory to provide self-supervising techniques that allow to quantify the reliability of the resulting projections.

After this mathematical and algorithmic preparation we will address the above mentioned two biomedical applications involving BSS: (a) artifact identification in Sec. 6 which demonstrates the separation quality that one can achieve for a MEG data set, where typical artifacts have been produced deliberately and (b) decomposition of event-related activity presented in Sec. 7. We conclude with a brief discussion.

2. The Blind Source Separation Model

In many signal processing applications one has only access to measurements of mixed, i.e. superimposed signals and the question is how to construct suitable projections that allow to demix and thus find the underlying (unmixed) signals of interest. Blind source separation (BSS) techniques aim exactly to reveal unknown sources using two ingredients (a) a model about the mixing process (typically a linear superposition cf. Eq. (1)) and (b) the assumption of statistical independence. As opposed to other signal processing techniques like beamforming (e.g. [Veen & Buckley, 1988]), BSS uses no geometrical information about the sensor array or the underlying sources, therefore BSS is called “blind”.

Let us assume that we measure a multivariate time-series and at time instant k, the observed n-dimensional data vector, $x(k) = [x_1(k), \ldots, x_n(k)]^T$ is given by the superposition

$$x(k) = \sum_{j=1}^{m} a_j s_j(k) = As(k). \quad (1)$$

As stated above the source signals, $s_1(k), \ldots, s_m(k)$, are unknown, along with the coefficients of the mixing process (matrix $A = [a_1, \ldots, a_m]$). The goal is therefore to estimate both unknowns from the measurements $x(k)$ and in principle all there is to do is to invert the mixing process

$$y(k) = \hat{s}(k) = Bx(k) = BAs(k), \quad (2)$$

where $B \approx A^{-1}$ is called the separating matrix.²

The general BSS problem requires $A$ to be an $n \times m$ matrix of full rank, with $n \geq m$ (i.e. there are at least as many mixtures as independent sources). In most algorithmic derivations, an equal number of sources and sensors is assumed.

In the model summarized by Eq. (1), and schematically illustrated in Fig. 1 we neglect

---

¹Note that due to the reviewing nature of this paper previously published material is used (cf. [Vigário, 1999; Vigário et al., 2000; Vigário & Oja, 2000; Meinecke et al., 2002; Vigário et al., 2002]).

²Unmixing is unique up to scaling and permutation, i.e. $BA = PD$, where $D$ is a diagonal matrix and $P$ is a permutation matrix.
any noise contribution. Some approaches to take additive noise into account can be found in
[Hyvärinen, 1999; Vigário, 1999; Müller et al., 1999; Kawanabe & Murata, 2000; Särelä et al., 2001; Hyvärinen et al., 2001].

Note that for decomposing brain data both the estimates of underlying source signals and the mixing matrix provide important information as the former reveals the temporal dynamics of a component (corresponding to a particular brain source), whereas the latter will give the corresponding spatial (stationary) field patterns, facilitating localization of the respective cortex activation.

3. Blind Source Separation Methods

A powerful method to perform BSS is independent component analysis (ICA). The literature on that subject is large and steadily growing (see Comon, 1994; Cardoso, 1998, 1999; Roberts & Everson, 2001; Hyvärinen et al., 2001) for reviews). Most of those methods try to estimate the mixing matrix $A$, respectively its inverse $B$, by optimizing an appropriate cost-function which measures the degree of statistical independence among the signal components.

A necessary condition for statistical independence is uncorrelatedness, therefore most algorithms enforce the data to be uncorrelated and unit-variance by a whitening (or sphering) transformation. This whitening step may also include a data compression which can be helpful to reduce/avoid overlearning (overfitting) effects that occur due to the limited number of sample points [Särelä & Vigário, 2001]. Whitening may be accomplished by PCA projection: $v(k) = Vx(k)$, with $E\{v(k)v(k)^T\} = I$. The whitening matrix $V$ is given by $V = \Lambda^{-1/2}U^T$, where $\Lambda = \text{diag}[\lambda(1), \ldots, \lambda(m)]$ is a diagonal matrix with the eigenvalues of the data covariance matrix $E\{x(k)x(k)^T\}$, and $U$ is a matrix with the corresponding eigenvectors as its columns. In terms of $v(k)$, the model (1) becomes $\hat{y}(k) = WAs(k)$, and it is clear that $W = VA$ is an orthogonal matrix, since otherwise the whitening condition would be violated [Hyvärinen et al., 2001]. Therefore, the solution is now sought in the form:

$$\hat{s}(k) = W^Tv(k). \quad (3)$$

Even if the whitening itself is not enough to get statistical independence (except for Gaussian signals), the problem has now been reduced to the determination of an orthogonal transformation (rotation). One possibility to get additional information to determine this “missing” rotation would be to take the higher-order statistics (HOS) into account and enforce independence by zeroing all higher-order cross-moments [Comon, 1994; Cardoso, 1999].

3.1. Exploiting higher-order statistics

3.1.1. Gradient-based techniques

Many popular algorithms for ICA are gradient-based techniques that optimize a contrast function which measures the statistical independence or the deviation from a Gaussian distribution — using the fact that a mixture of independent non-Gaussian sources becomes more Gaussian.

For example the FastICA algorithm [Hyvärinen & Oja, 1997] uses the fourth-order cumulant — also called the kurtosis — as a measure of non-Gaussianity. The kurtosis of the $i$th source signal is defined as $\kurt(s_i) = E\{s_i^4\} - 3\{E\{s_i^2\}\}^2$, where $E\{\cdot\}$ denotes the mathematical expectation value of the bracketed quantity. The kurtosis is negative for source signals whose amplitudes have sub-Gaussian probability densities (distributions flatter than Gaussian), positive for super-Gaussian (sharper than Gaussian, and with longer tails), and zero for Gaussian densities. Maximizing the norm of the kurtosis leads to the identification of non-Gaussian sources. Consider a linear combination $y = w^Tv$ of the white random vector $v$, with $\|w\| = 1$. Then $E\{y^2\} = 1$ and $\kurt(y) = E\{y^4\} - 3$. Then the gradient of the kurtosis with respect to $w$ is $4E\{v(w^Tv)^3\}$.

FastICA is a deflation algorithm that minimizes the kurtosis by finding one of the columns of the separating matrix $W$ (noted $w$) at a time and thus
identifying the respective independent source. In the 1th iteration FastICA computes
\[
W_i^* = E\{v(W_{i-1}^T)^3\} - 3W_{i-1}
\]
\[
W_i = W_i^*/||W_i^*||.
\]
In order to estimate up to \(m\) components, the algorithm is run repeatedly removing the information contained in components that were already extracted (for matlab code see http://www.cis.hut.fi/projects/ica/fastica/ and cf. [Hyvärinen & Oja, 1997; Vigário, 1999; Hyvärinen et al., 2001] for further details).

In general, gradient-based ICA algorithms consider cost functions of the form \(J(B) = E\{G(y)\}\) where \(G\) is some nonlinear function and determine the solution by iterative updates of \(B\) using
\[
\frac{\partial J(B)}{\partial B} = E\{g(y)x^T\},
\]
where \(g(y)\) is the gradient of \(G(y)\) and \(y = Bx\). Assuming further that \(B\) is invertible, we know that \(x = B^{-1}y\) and obtain
\[
\frac{\partial J(B)}{\partial B} = E\{g(y)y^T\}(B^T)^{-1}.
\]
Provided that the nonlinear functions \(G(y)\) are suitably chosen, the algorithm converges to a stationary point, where the components of \(y\) are mutually independent [Amari et al., 1996].

A Matlab\textsuperscript{TM} package that implements algorithms of this type (e.g. INFOMAX [Bell & Sejnowski, 1995]) can be found at:
http://www.sccn.ucsd.edu/~scott/ica-download-form.html

3.1.2. JADE

In contrast to these gradient-based techniques there are also algebraic methods for BSS. For example, Cardoso et al. developed the JADE [Cardoso & Souloumiac, 1993] algorithm based on the joint diagonalization of matrices. After whitening and possible dimension reduction, a set of matrices obtained from “parallel slices” (Eigen matrices) of the fourth-order cumulant tensor is approximately diagonalized with a single orthogonal transformation [Cardoso & Souloumiac, 1993; Comon, 1994]. Usually JADE provides excellent results on low dimensional data if sufficiently many sample points are available. However, for high dimensional problems like MEG the effort for storing and processing the fourth-order cumulants is \(O(m^4)\) which may be prohibitively heavy. Therefore for such applications JADE needs to project down to a lower dimensional subspace. A Matlab\textsuperscript{TM} package that performs the JADE can be found in the address:
ftp://sig.enst.fr/pub/jfc/Algo/Jade

3.2. Exploiting time structure

In the above ICA approaches to BSS the signals were assumed to be identically independently distributed (i.i.d.). However, many “natural” signals, like speech signals or in particular, EEG and MEG signals are characterized by their rich dynamical time structure. Therefore, it seems natural to directly exploit temporal, resp. spectral information, for blind source separation.

An attractive algorithmic framework to deal with this problem is — similar to the JADE case above — the simultaneous diagonalization of appropriately defined matrices by algebraic methods (cf. [Tong et al., 1991; Molgedey & Schuster, 1994; Belouchrani et al., 1997; Pham & Garat, 1997; Ziehe & Müller, 1998; Wu & Principe, 1999; Ziehe et al., 2000b]).

In particular, the TDSEP algorithm [Ziehe & Müller, 1998] finds an estimate of the mixing matrix \(A\) by simultaneously diagonalizing a set of several symmetrized time-lagged correlation matrices defined as
\[
C(x, \tau) = \frac{1}{2T} \sum (x(t)x^T(t-\tau) + x(t-\tau)x^T(t))
\]
Under the assumption that the source signals \(s(t)\) are mutually uncorrelated (over time) and possess a pronounced auto-correlation [Ziehe & Müller, 1998], this solves the BSS problem since the covariance matrices of the source signals are transformed according to
\[
C(x, \tau) = AC(s, \tau)A^T.
\]
This algorithm can also be seen as an efficient way to minimize the cost function
\[
J(B) = \sum_{\tau} \sum_{i \neq j} (C_{ij}(Bx, \tau))^2 ,
\]
that measures the cross-correlation of the outputs \(y = Bx\) for several time-lags \(\tau\). Note that though in theory the quality of source separation depends on\footnote{The corresponding independent source signal can then be found using Eq. (3).}
the selection of the lag parameters $\tau$ which should ideally be adapted to the auto-correlation structure of the signals; in practice, however, it turns out that TDSEP is quite robust w.r.t. the choice of these parameters, as long as sufficiently many matrices are used for simultaneous diagonalization (for details see [Müller et al., 1999]).

As with the ICA methods presented in Sec. 3.1.1, the simultaneous diagonalization procedure that is employed here performs first a sphering or whitening step followed by a rotation. The determination of the rotation matrix $W$ is achieved by the extended Jacobi-type method by Cardoso and Comon [1996], applied to the set of time-delayed correlation matrices. This is in analogy to the simultaneous diagonalization of eigenmatrices of the cumulant tensor in JADE and in fact both algorithms can be combined in a straightforward manner and by this enlarging the set of matrices to be diagonalized (cf. [Müller et al., 1999]). From the computational point of view TDSEP is particularly well suited for the processing of high-dimensional multichannel data typically encountered in physiological recordings. A Matlab implementation of this algorithm is available (see http://www.first.fhg.de/~ziehe/research.html).

4. Limits and Problems of Source Separation Algorithms

4.1. How good is the underlying model assumption

Although BSS uses a “weak” model it must be emphasized that it can only yield results within its model class. This is the same for all unsupervised learning algorithms: if we run PCA (see e.g. [Duda et al., 2001]) then all we can expect is an orthogonal decomposition result that describes the data optimal in the least squares sense (i.e. orthogonal directions of the data cloud with maximal variance or energy are found). This however does not mean that PCA provides a good basis for understanding the data, if say a nonorthogonal transformation into an ICA basis would have been more appropriate to describe the intrinsic nature of the data.\[4\]

It is instructive to explain the difference between PCA and ICA in Fig. 2, which show toy data constructed to make this difference clear. In Fig. 2(a) we see a scatterplot of the data points given by two mixed super-Gaussian sources. The projection of the data points onto the coordinate axes yields the observed signals. It can easily be seen that the two signals are not independent. Figure 2(b) shows this data set in the PCA basis (i.e. the eigenvector basis of the covariance matrix $\mathbf{C(x)}$). Note, that PCA results in an orthogonal transformation, therefore Fig. 2(b) is a rotated version of Fig. 2(a). Clearly PCA does not provide a good basis for describing the data, since the signals are still mixed.

Figures 2(c) and 2(d) show the two steps of source separation by JADE: first a whitening step rescales the data set such that the covariance matrix is the identity. The signals are now uncorrelated but not independent. A rotation then transforms the data set into the ICA basis. It is easy to see, that this basis provides an excellent basis to further understand the given data.

The efficiency of high-order statistics to produce good ICA algorithms (the kurtosis based FastICA, JADE, . . .) comes with a price: a sensitivity to outliers.\[5\] In fact, these often turn out to be the leading factors in the search for the underlying sources.

Another limitation of the presented BSS approaches is that one can extract only up to as many underlying sources as the number of sensors. In the brain, however, the multitude of microscopic sources outnumbers by far the number of sensors. Nevertheless, it can be argued that the total number of macroscopically observable sources, active at a given time, is of a much smaller count.

4.2. The necessary amount of data: Overfitting

When the number of samples of the observed data is insufficient to explain the high-dimensional independent source signals, we may encounter over-learning or overfitting effects in BSS algorithms. This result is true for practically any type of contrast-based linear ICA implementation [Särelä & Vigário, 2001]. In the extreme case of an equal

\[4\] The same argument holds of course for nonlinear decompositions like kernel PCA [Schölkopf et al., 1998].

\[5\] PCA is also nonrobust to outliers but the use of robust methods allows to obtain less sensitive PCA algorithms, e.g. [Huber, 1981].
number of independent source signals, mixtures and sample sizes, it is shown that the optimal solution is the one that produces source signals that are zero almost everywhere except for a single spike or bump. In the referred communication, experimental evidence of this overfitting effect is given, using both simulated and real MEG data.

Temporal decorrelation algorithms, such as the TDSEP, may suffer from a different type of overlearning, yielding a series of sinusoidal components of various frequency contents.

Furthermore, channel noise, explicitly ignored in the model of Eq. (1), effectively doubles the number of independent sources, rendering the overlearning problem ever more present. In some applications, we may construct an approximate noise model. Projections to signal spaces orthogonal to the noise space can then be performed [Hyvärinen, 1998; Müller et al., 1999; Hyvärinen et al., 2001; Särelä et al., 2001].

When utilizing algorithms requiring pre-whitening, this preparatory stage can be used to reduce the dimensionality of the data. The compression reduces the overlearning effect, especially if the rejected components contain mainly noise (i.e. unnecessary information).

Ultimately, a good reliability measure of the components found may help to determine whether we are in presence of clear overlearning, or of an overlearning-like “true” source. Furthermore, such measures could help determine whether the model is appropriate at all or which model, i.e. BSS algorithm is to be chosen. The first such measure that computes the stability under bootstrap resampling was recently proposed by Meinecke et al. [2002] and will be discussed in more detail in the next section.

5. Reliability of ICA Projections

5.1. The Bootstrap resampling

The reliability of ICA projections can be estimated by resampling techniques. Resampling is a statistical method which gives e.g. the variance of estimators only from one set of data \( \mathbf{x} \) at hand by virtue of modern computer power. Among such procedures, the Jackknife and the Bootstrap are most well-known (see e.g. [Efron & Tibshirani, 1993; Shao & Tu, 1995]). The Jackknife produces surrogate data sets by deleting only one datum each time from the original data set. There are generalizations of this approach like the delete-\( k \) Jackknife which deletes more than one datum at a time. The bootstrap is a more general approach and has been widely used in data analysis recently.

Let us consider the standard case where a sample \( x_1, \ldots, x_T \) with i.i.d. observations from a distribution \( F \) is given. We will write the data by the vector \( \mathbf{x} = (x_1, \ldots, x_T) \). A scalar parameter \( \theta \) is estimated with an estimator \( \hat{\theta}_i(\mathbf{x}) \). To get some information about the accuracy of this estimate we build \( B \) bootstrap samples \( \mathbf{x}^{vb} = (x_1^{vb}, \ldots, x_T^{vb}) \) with \( b = 1, \ldots, B \) by drawing \( T \) data points with replacement from the given sample. We remark that some data points might occur several times, while others do not occur at all in a particular bootstrap sample. On each surrogate \( \mathbf{x}^{vb} \), the estimator \( \hat{\theta}^{vb}_i = \hat{\theta}_i(\mathbf{x}^{vb}) \) is calculated,
so we have $B$ estimators $\hat{\theta}_i^1, \ldots, \hat{\theta}_i^B$. We can now for example estimate the variance of $\hat{\theta}_i$ by calculating the variance of the bootstrap estimators $\hat{\theta}_i^*$.

There is a wide literature on statistical properties of the Bootstrap and its extensions, which supports the use of resampling procedures. For example, it can be shown that for i.i.d. data the bootstrap estimators of the distributions of many commonly used statistics are consistent [Shao & Tu, 1995].

### 5.2. Resampling for BSS/ICA

In the ICA-case the parameters to be estimated are the one-dimensional source signal subspaces (since the mixing matrix is defined only up to a free scale and permutation) and a natural distance measure between two such subspaces is the angle difference between them.

It is straightforward to apply bootstrap resampling for i.i.d. data and for algorithms that do not use temporal structure. Less obvious is to construct a time structure preserving resampling. A simple bootstrap approach would clearly destroy temporal structure, but it can be generalized such that methods like TDSEP that use temporal correlations are still applicable. Consider a time series of length $T$. The bootstrap resampling defines a series $\{\tilde{x}_t\}$ with $\sum a_t = T$ and each $a_t$ tells how often the data point $x(t_0)$ has been drawn. Using this, we can calculate the resampled time-lagged correlation matrices as

$$\hat{C}^{*\tau}_{ij} = \frac{1}{2T} \sum_{t} a_t \cdot \{x_i(t)x_j(t-\tau) + x_i(t-\tau)x_j(t)\}.$$  (6)

The crucial idea on how to find stable higher dimensional source signal subspaces is to calculate not the overall rotation for each projection direction, but to decompose each rotation into $N(N-1)/2$ elementary rotations within all two-dimensional planes spanned by the coordinate axes. This can be carried out by taking the matrix logarithm

$$\hat{\alpha}^{eb} = \ln(\hat{R}^{eb}).$$  (7)

Here, each component $\alpha_{ij}$ of $\alpha$ is the angle of a rotation in the $i$–$j$-plane (see Appendix).

If we now calculate the variance of the $\alpha$-matrix component-wise, we obtain a separability matrix

$$\hat{S}_{ij} = \sqrt{\frac{1}{B} \sum_{b=1}^{B} (\hat{\alpha}_{ij}^{eb})^2}.$$  (8)

The value $\hat{S}_{ij}$ of this matrix measures how unstable the BSS solution is with respect to a rotation in the $i$–$j$-plane, i.e. how reliable the respective one-dimensional subspaces $i$ and $j$ can be separated from each other. Note that a low $\hat{S}_{ij}$ corresponds to good separability.

If the used BSS algorithm was successful in separating the independent subspaces, the separability matrix should have a block structure that groups together one-dimensional ICA projections that cannot be reliably separated from each other but form a multidimensional stable subspace. Therefore this method allows to reveal structural dependencies which provides highly important and relevant information for subsequent biomedical interpretation.

### 6. Artifact Contaminated MEG Data

#### 6.1. Description of the data

As a first example, we illustrate the efficient analysis of artifact contaminated MEG data with BSS
algorithms. The data under study has been recorded at the Brain Research Unit of the Low Temperature Laboratory of the Helsinki University of Technology, in a magnetically shielded room using a 122-channel whole-scalp Neuromag-122 SQUID magnetometer [Hämäläinen et al., 1993].

The measured subject was asked to deliberately blink and make horizontal saccades, in order to produce typical ocular artifacts. Moreover, to produce myographic artifacts, the subject was asked to bite his teeth for as long as 20 seconds. Yet another artifact was created by placing a digital watch one meter away from the helmet into the shielded room. Further information on the data can be found in [Vigario et al., 1998].

Figure 3 presents a sample of the measured signals, with a duration of approximately one minute (i.e. 8000 samples), showing brain activity blurred by the aforementioned artifacts.

6.2. Results

In earlier studies, FastICA as well as Topographic ICA have been applied to the data and results can be found in [Vigário et al., 1998; Hyvärinen et al., 2001]. In the experiments reported here, we used the source separation algorithms JADE and TDSEP (with $\tau = \{0, 1, \ldots, 20\}$). In order to mitigate the overlearning effect, we applied the BSS not to all 122 signals, but to the first 20 principal components. Furthermore, following the recommendation in [Särelä & Vigário, 2001], the data was high-pass filtered with a cut-off frequency of 0.5 Hz (using again the same setting as in [Hyvärinen et al., 2001]).
Fig. 4. JADE decomposition of the artifact time series. 11: horizontal eye movements, 12: vertical eye movements (blink), 13: cardiac cycle, 14–20: muscle activity (biting).

Figure 4 shows the JADE decomposition. The separability matrix, on the right, highlights the existence of three reliable one-dimensional independent components. The temporal structure of these components allows us to give them clear physical interpretations: while signals 11 and 12 capture the horizontal and vertical eye movements (or blinks), component 13 reflects mainly the cardiac contamination present in the MEG recordings.

The block of independent components 14–20 corresponds to muscle activity. Hence, it can be said that the muscle artifacts need, at least, a six-dimensional space to fully explain their contribution in this data set. These results agree and corroborate
the findings of topographic ICA [Hyvärinen et al., 2001].

The signal of the digital watch, completely invisible in the raw data, now appears concentrated in component 5.

Figure 5 shows the front, left, top and right views of the three one-dimensional artifact components found by the JADE algorithm (cf. Fig. 4). The top set corresponds to the horizontal eye activity. The middle one is characteristic for the vertical and blink eye artifacts, whereas the last row agrees with the expected pattern of the cardiac signal [Jousmäki & Hari, 1996; Vigário et al., 1998].

In Fig. 6 we see the TDSEP decomposition of the same data. Here the separability matrix reveals that TDSEP is able to find more independent

![Fig. 6. TDSEP decomposition of the artifact time series. 1: horizontal eye movements, 2: vertical eye movements (blink), 3: watch, 12: cardiac cycle, 15-20: muscle activity.](image)

![Fig. 7. (a) The first two principal components display eye artifacts. (b) A better representation of these artifacts is found in an ICA basis (JADE).](image)
components with one-dimensional reliability than JADE. Among the estimated source signals with clear interpretations are both eye artifacts (Nos. 1 and 2), the digital watch (No. 3) and the cardiac cycle (No. 12). Note that, once again, the muscle activity is grouped as a six-dimensional block of components 14–20.

Because of the strong historical and methodological relations that exist between PCA and ICA, an exemplary comparison of the separations achieved by such techniques is shown in Fig. 7. The upper panel ((a) and (b)) shows the time courses and a scatterplot of the strongest two principal components. Coincidently, they capture mainly the eye movement artifacts. For a good comparison, the lower panel ((c) and (d)) displays the time courses and scatterplot of the corresponding independent components, found by JADE. A very significant improvement of the signal-to-noise ratio is observed in the ICA decomposition. Several blinks became now visible, which were buried under muscle activity and other sources of noise. Although not shown here, these denoising/separation differences are even more prominent for the weaker signals of the mixture (e.g. the watch): Here PCA fails completely in finding them, since it looks only at the power of the signals, while ICA is able to find reasonable good source reconstruction.

7. Analysis of Multimodal Evoked Fields

In a second example, we analyze event-related MEG recordings, averaged over many trials. While this is needed to improve the signal-to-noise ratio, the effective number of data points available for BSS is considerably decreased. Therefore, this dataset is suited to study overlearning behavior that can occur when one imprudently applies the BSS methods presented in this paper. Furthermore, we will show how the unreliable nature of these behaviors can be used to our advantage. The ultimate goal is again to derive good strategies to identify interpretable components of brain responses.

7.1. Description of the data

The data consists of averaged responses of a single subject to vibrotactile stimulation, generated with a bass-reflex loudspeaker and a tube delivering the tone into the magnetically shielded room. The tube was coupled to a balloon which was held by the subject with both hands. Both tactile and the concomitant auditory stimuli were thus present (further details on the stimuli and the setup can be found in [Jousmäki & Hari, 1999]; an application of ICA to this data can be found in [Vigário et al., 2000], and references therein).

Figure 8 shows the complete set of 122 averaged evoked responses. The two inserts in the figure indicate roughly the shape of signals of interest present in the measurements. The left one shows a sharp and early response, originated in the primary somatosensory cortex while the right insert has a broader signal, with longer latency than the previous one, which is associated with the primary auditory responses.

7.2. Results

The first results we show are left here as a “word of warning” against overlearning. In fact, we used the same strategy for this data as in the previous section, reducing the dimension from 122 to 20. Yet, the amount of available samples is now much smaller. Therefore, strong overlearning is produced by both JADE and TDSEP (see Figs. 9 and 10). Higher-order statistics methods, such as JADE, typically produce bumpy components, whereas the overlearning regime of TDSEP is characterized by sinusodial source estimates [Särelä & Vigário, 2001; Ziehe & Müller, 1998]. Note that nearly no overlearned component show a good degree of reliability. The exceptions are the 13th JADE component, as well as the last three TDSEP components. The later ones, with a frequency content of 100 Hz and 50 Hz, are in fact reasonable components for this study, as they take into account the powerline artifacts, present on the whole data. For the JADE overlearn, it will become clear, from the subsequent analysis, that it is a simplified version of a reliable brain response to the vibrotactile stimulation.

A more reasonable data compression is achieved when retaining only the strongest seven principal components. Figure 11 shows the result of the application of JADE to such compressed data. Here we see that only the first three components are clearly reliable. The first two, with evident blocking, peak around 100 ms after the onset of the vibrotactile stimulation, and are therefore expected to be mainly the response of the brain to the auditory stimulus. The third has a much shorter duration...
Original Averages

Fig. 8. Averaged responses to vibrotactile stimuli measured with a 122 channel MEG system (from [Vigario et al., 2000]).

Fig. 9. TDSEP decomposition of the vibrotactile data in an overlearning regime.
and latency, very typical for somatosensory evoked responses (for further considerations on these matters, please refer to [Jousmäki & Hari, 1999; Vigário & Oja, 2000]).

Figures 12 and 13 show the mixing vectors corresponding to the first two JADE ICs, disposed as three-dimensional field patterns. Several views of these patterns are supplied to simplify its...
inspection. Furthermore, the localization of equivalent current dipoles (ECDs), explaining best the aforementioned patterns, is marked over MRI scans of the measured subject. It is clearly visible that those dipoles exist within the sylvian fissure, as should be expected for such early auditory brain responses.

Figure 14 presents similar plottings as the previous ones, now for the third component. Note that the field patterns have a very clear dipolar structure, and that the ECDs that best fit these patterns fall on the expected primary somatosensory cortex area. Once again, these results agree with the ones observed in [Jousmäki & Hari, 1999; Vigário & Oja, 2000].

According to our prior knowledge we expect, that the components of interest in this vibrotactile evoked response study are temporal and very focal, which results in a high sparsity. Furthermore, due to the averaging, the number of data point becomes very small and almost all periodicity of the data is lost. In such cases we anticipate that HOS based methods like JADE are much better qualified to separate the signals than temporal decorrelation methods like TDSEP.

In fact, Fig. 15 shows that the only somewhat reliable components are the 6th and the 7th, which could be seen as depicting trends. The signals of interest, coincidently still the first three, are not clearly separated from each other, as can be seen from the blocking effect of the separability matrix. Furthermore, it seems evident that the first TDSEP component, in an attempt to produce its preferred periodic solutions, accounts for most of the information contained in the first and third JADE components, i.e. the somatosensory and some of the auditory responses.

The field patterns of Fig. 16 clearly show contributions from both signals, i.e. the left and top...
Fig. 13. Field patterns of the second JADE component. Mainly the right-hand side component of the auditory brain response.

Fig. 14. Field patterns of the third JADE component, corresponding to the somatosensory response.
views show a pattern that is clearly related to the somatosensory response of the left hemisphere. Yet, the very strong auditory response, illustrated on the right view of the field pattern, overshadows the right somatosensory response. Clearly the reliability analysis would suggest to use JADE over TDSEP for this particular application.

8. Concluding Remarks

What makes blind source separation an appealing method for the analysis of neurobiological data, is that it uses a “weak” model, i.e. statistical independence (e.g. JADE) or independence over time (e.g. TDSEP) respectively. Although no strong model (such as the ones based on physiological models) is imposed on the data, note that ICA algorithms still extract components, which have very good neurophysiological plausibility.

So far, we could count on a few measures to assess the good quality of our BSS decomposition of neurobiological recordings, such as the inspection of the time/frequency courses of the components and their respective field patterns. Now, from a signal processing viewpoint, we introduce an additional unsupervised evaluation method for the goodness of the analyses: the resampling-based reliability measure. The basis for this assessment is the concept of stability: small distortions due to resampling should not change the projections found. If they change, then this is a good indicator that the underlying assumption of the BSS algorithm has failed.
In the reported experiments, we have used raw and averaged biomagnetic brain signals and demonstrated the merits and pitfalls of different BSS approaches. Special emphasis was given to the validation of the ICA model. Interestingly, we find that the reliability method can detect solutions that are due to overfitting, and thus allows algorithm and even hyperparameter selection in the unsupervised learning scenario of BSS.

ICA was able to isolate artifacts in MEG. In the event-related study, ICA differentiated efficiently between somatosensory and auditory brain responses in the case of combined auditory and vibrotactile stimulation. In addition, the independent components, found with no other modeling assumption than their statistical independence (or independence over time), exhibit field patterns in good agreement with the conventional current dipole models. The equivalent current dipoles corresponding to the independent components are located in brain regions expected to be activated by the respective stimuli.

Future research will be devoted to incorporating prior knowledge into BSS models. In some preliminary studies, we have observed that, as we depart from purely statistically based assumptions, we might get closer to physiologically more plausible decompositions of electromagnetic brain signals [Vigário, 2000; Vigário et al., 2000; Särelä et al., 2001]. On the other hand, fitting neural sources in a classical framework, may be hard if some temporal overlap is present in their activations. A well balanced use of both the model and appropriate priors is therefore needed in order to ultimately obtain a good exploratory decomposition technique for the application at hand.

Acknowledgments

Valuable discussions with M. Kawanabe, C. Schäfer, J. Kohlmorgen, G. Nolte, B. Blankertz, G. Curio, J. Särelä, Harri Valpola and participants of the Potsdam workshop on ERPs 2001 are gratefully acknowledged. The authors thank the Low Temperature Laboratory at HUTs for the permission to use their data in this work. R. Vigário was funded by EU (Marie Curie fellowship HPMF-CT-2000-00813). K.-R. Müller and A. Ziehe were partly funded by EU (IST-1999-14190 – BLISS) and the BMBF in the BCI project.

References


Appendix A
Rotation Angles and Rotation Matrices

The rotation matrices in $N$-dimensional space form a representation of the special orthogonal group $SO(N)$. This means, that for all $R \in SO(N)$

\[ RR^T = 1 \tag{A.1} \]
\[ \det(R) = +1 \tag{A.2} \]

Any orthogonal matrix $R$ can be written as the exponential of a single antisymmetric matrix $\alpha$

\[ R = e^{\alpha} = \sum_{n=0}^{\infty} \frac{1}{n!}\alpha^n \]

To see this, just transpose $R = e^{\alpha^T}$:

\[ R^T = (e^{\alpha})^T = e^{\alpha^T} = e^{-\alpha} = R^{-1} \]

A basis $\{M^{ij}\}$ in the space of all antisymmetric matrices can be defined by

\[ (M^{ij})_{ab} = \delta_{a}^{i}\delta_{b}^{j} - \delta_{a}^{j}\delta_{b}^{i} \tag{A.3} \]

where $\delta_{i}^{a}$ denotes the Kronecker-Symbol: $\delta_{i}^{a} = 1$ if $i = a$ and $\delta_{i}^{a} = 0$ otherwise. Note, that each $M^{ij}$ is a $N$ by $N$ matrix, $a$ and $b$ specify the matrix entries, i.e. $a$ is the row- and $b$ the column-index. Using this, every orthogonal matrix can be written as

\[ R = \exp \left( \frac{1}{2} \sum_{i,j=1}^{N} \alpha_{ij} M^{ij} \right) \tag{A.4} \]

Due to the antisymmetry of the $M^{ij}$, one can choose the $\alpha_{ij}$ to be antisymmetric, too, without loss of generality. If we do so, each $\alpha_{ij}$ corresponds to the angle of a basic rotation within the $i$–$j$-plane. More precisely, $\exp(\alpha_{ij}M^{ij})$ is an orthogonal transformation that rotates the $i$-axis towards the $j$-axis by the angle $\alpha_{ij}$.

So, given an arbitrary rotation matrix $R$, it is always straightforward to decompose it into elementary rotations within different planes by taking the matrix-logarithm:

\[ \alpha = \ln(R) \tag{A.5} \]