OFI: OPTIMAL FILTERING ALGORITHMS FOR SOURCE SEPARATION

Andreas Ziehe[†], *Guido Nolte*[‡], *Gabriel Curio*[‡] *and Klaus-Robert Müller*[†*]

[†]GMD FIRST.IDA, Kekuléstr. 7, 12489 Berlin, Germany
[*]University of Potsdam, Am Neuen Palais 10, 14469 Potsdam, Germany
[‡]Neurophysics Group, UKBF, FU Berlin, Hindenburgdamm 30, 12200 Berlin, Germany

ABSTRACT

We analyze second-order methods for signal separation and show how specific filtering operators can be used to improve the separation performance. It is the goal to exploit all spectral information in an optimal way. A perturbation theoretical approach is used to derive an explicit solution for the 2×2 mixing case and an approximate solution for the $M \times M$ case is given. The usefulness of our optimal filtering method (OFI) is demonstrated by simulations.

1. INTRODUCTION

Blind source separation (BSS) methods have been successfully applied for a variety of problems (see e.g. [2, 10, 3, 6, 12, 13, 18, 19, 20]). Two BSS approaches are distinguished: those relying on higher-order statistics, and second order algorithms that exploit spectral information. Recently also hybrids have been proposed [14]. Since most naturally occurring signals carry spectral information we will concentrate on second-order algorithms in this paper.

We propose to view second order source separation in terms of filtering theory inspired by Wiener filtering and prove an optimal filter solution with a perturbation theoretical approach. It turns out that for special conditions (no noise, no spatial pre-whitening) our OFI algorithm is equivalent to the maximum likelihood solution obtained by Pham et al. [16]. We pursue two goals in our paper: (1) we would like to increase the general understanding of why second order BSS algorithms work well and (2) we give a second order algorithm based on the idea of optimal filtering. The filtering approach allows us to make optimal use of differences in spectral information between the source signals.

First we briefly state the source separation problem. Consider M unknown sources that generate M statistically independent time series $s_i(t)$ $i = 1, ..., M$, $t = 1, ..., T$ that are spatially uncorrelated but have a non-delta temporal autocorrelation function. A sensor array consisting of M sensors $x_j(t)$ measures a stationary linear superposition $x_j(t) = \sum_i W_{ji} s_i(t)$ ($\vec{x}(t) = W\vec{s}(t)$). The goal is to identify W in this model and hence to blindly reconstruct $\vec{s}(t)$ given only $\vec{x}(t)$. It is well-known that in the case of *temporally correlated sources* the linear BSS problem can be solved by a simultaneous (approximate) diagonalization of two (or more) generalized covariance matrices [4, 13, 19]. One makes use of temporal information in the signals by constructing (e.g. [13, 17, 4, 19])

$$C_{ij}^{\tau} := \frac{1}{2} \sum_t x_i(t)(x_j(t-\tau) + x_j(t+\tau)) \qquad (1)$$

for $\tau = 0$ and $\tau = \tau_0 \neq 0$, and by finding a matrix U which simultaneously diagonalizes C^{τ} for both values of τ. Then U is an estimate for W (up to scaling and permutation). A solution always exists as long as the covariance matrices are symmetric (general eigenvalue problem). Though intuitively simple, it is so far not understood which values τ_0 should be chosen and therefore the performance of the algorithm can be rather poor for a wrong choice. All algorithms so far – even the ones making use of several covariance matrices (cf. [19, 4]) – rely on heuristics for choosing the appropriate delay structure of the covariance matrices and are therefore susceptible to the discussed problem.

In the following section we will first derive the theoretical framework. Then we describe a new practical algorithm and present some simulation results and finally give a conclusion.

2. OPTIMAL SEPARATING FILTERS

We will now address the question of the optimal use of spectral information. For the moment we assume that we have *all* information about the original source spectra. For the time being we will consider two signals, however the generalization to M signals is in principle

A.Z. was partly funded by DFG under contracts JA 379/51 and JA 379/71. G.N. and G.C. were supported by DFG grant MA 1782/3-1.

straight forward though not all equations can still be solved as nicely as below. As a first step we generalize the operation of time-shifts in Eq. (1) to allow for arbitrary linear transformations in time. Then the most general symmetric covariance matrices read

$$C_{ij}^{a,b} := \frac{1}{2}\sum_t x_i(t)(\phi^{a,b} \star x_j)(t) + i \leftrightarrow j, \qquad (2)$$

where $\star$ denotes convolution and ϕ^a and ϕ^b are two arbitrary filter functions that filter x. Note that a filtering of both signals $x_i(t)$ and $x_j(t)$ can always be written in the form (2). In fact, existing blind source separation methods (like [13, 19, 11]) utilized 'implicit' filtering where so far the filters were set heuristically. It is interesting to note that this choice for the matrices $C^{a,b}$ also belongs to the class of admissible estimating functions introduced by Amari [1]. The standard operation of symmetrized time-shifts corresponds in Fourier space to the filter $\Phi^\tau(k) \sim \cos(2\pi\tau k/T)$ which directly follows from writing (1) in the form (2) and then Fourier transforming $\phi(t)$. Taking this point of view it becomes apparent that second-order signal separation methods like the TDSEP algorithm [19] implicitly use a set (or basis) of cosine functions with frequencies that are determined by the delay values chosen for the respective covariance matrices. Those algorithms work well since within this basis sophisticated filters can be constructed efficiently.

From now we will entirely work in Fourier space and denote the Fourier transformed filters and signals by capital letters, and – in order to construct symmetric covariance matrices – we will set all filters to be real valued throughout the paper.

Optimal Filtering With Prewhitening Before coming to the general construction of the two filters Φ^a and Φ^b we discuss a simpler case by considering spatially whitened signals, i.e. we set $\Phi^a(k) = 1$, and scale the signals such that $\sum_k |S_i(k)|^2 = 1$.

We take a *perturbation theoretical* approach. For this we expand the eigenvectors $\vec{u}_i$ of the estimated mixing matrix U to first order in the basis of the true eigenvectors $\vec{w}_i$ (e.g. $\vec{u}_1 = \vec{w}_1 + \alpha\vec{w}_2$)

$$\vec{u}_i = \vec{w}_i + \Delta\vec{w}_i \approx \vec{w}_i + \sum_{j\neq i} \frac{\Delta C_{ij}}{\lambda_j - \lambda_i}\vec{w}_j, \qquad (3)$$

where $\lambda_i = 2\sum_k \Phi(k)S_i^2(k)$ are the eigenvalues of the true (unknown) covariance matrix of the original sources C. To find an optimal filter function Φ we minimize the expected value[1] of α^2. In two dimensions this means minimizing the functional

$$L(\Phi) := \overline{\alpha^2} = \frac{\overline{(\Delta C_{12})^2}}{(\lambda_1 - \lambda_2)^2}. \qquad (4)$$

For the calculation of the expected value we assume independent and stationary signals

$$\overline{S_i(k)^* S_j(k')} = P_i^2(k)\delta_{kk'}\delta_{ij}$$

with $P_i^2(k) := \overline{|S_i(k)|^2}$ (the spectrum of the $i.th$ independent signal). Inserting this into (4) leads to the result that for white signals the misestimation of the source separation reads for $T >> 1$

$$L(\Phi) = \frac{\sum_k \Phi^2(k)H(k)}{\left(\sum_k \Phi(k)(P_1^2(k) - P_2^2(k))\right)^2} \qquad (5)$$

with the 'weighting function'

$$H(k) = P_1^2(k)P_2^2(k). \qquad (6)$$

The derivative of $L(\Phi)$ with respect to $\Phi(k')$ yields the optimal filter

$$\Phi_{opt}(k) = \frac{P_1^2(k) - P_2^2(k)}{H(k)} \qquad (7)$$

after getting rid of an arbitrary factor. Let us discuss this solution:
(a) The (local) sign of the filters matters. Without explicit derivation we merely state that if we would have used a quadratic form ($\Phi \to \Phi^2$), corresponding to an ansatz of filtering both x_i and x_j in Eq.(2), the respective optimal filters (basically (7) with negative values set to zero) would have been worse by about a factor 2.
(b) The filter diverges if e.g. for some frequency k, the power spectrum $P_1^2(k) = 0$ and $P_2^2(k) \neq 0$. This corresponds to the trivial separation when for some k the first signal is absent. Filtering out this component gives an exact solution.
(c) For large and white channel noise, $H(k)$ can be set to a constant and in this limit the optimal filter is essentially the Fourier transform of the difference of the autocorrelation functions. Expressing the optimal filter as a superposition of the time-shift-filters is equivalent to inverse Fourier transforming the optimal filter, and hence the τ_0 which leads to the *best* approximation of the optimal filter is the one at which the *difference* of the autocorrelation functions of the respective sources is *maximal*.

To illustrate the optimal filter method let us consider two sources with their respective power spectra $|P_1|^2$ and $|P_2|^2$ (cf. Fig. 1). Eq.(7) gives the formula

[1] The expectation is denoted by overlining.

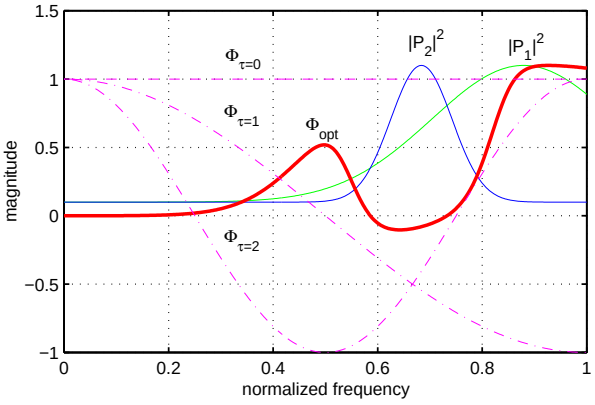

Figure 1: Source spectra and filters. $\Phi_{\tau=0,1,2}$ are the filters implicitly used by [13]; Φ_{opt} denotes the optimal filter according to Eq.(7).

for computing the optimal filter denoted by Φ_{opt} in the plot. The filter characteristics of the Molgedey-Schuster (MS) algorithm [13] using the covariance matrices C_τ (for $\tau = 0, 1, 2$ respectively) are shown for comparison. $C_{\tau=2}$ gives the best performance for single delays. We can clearly see in Fig. 1 that the optimal filter captures the differences of the spectra much better than the filters implied by the Molgedey-Schuster algorithm.

Optimal Filtering Without Prewhitening Let us now come to the general case with two different filters. Similar to the previous case simultaneous diagonalization of the covariance matrices leads to an estimate $U = (\vec{u}_1, \vec{u}_2)$ which can be expressed in the basis of the true mixing matrix $W = (\vec{w}_1, \vec{w}_2)$ as $\vec{u}_1 = \vec{w}_1 + \alpha_{21}\vec{w}_2$ and $\vec{u}_2 = \vec{w}_2 + \alpha_{12}\vec{w}_1$ with α_{21} and α_{12} being stochastic variables. The optimal filters are defined to be the ones which minimize $\overline{\alpha_{21}^2} + \overline{\alpha_{12}^2}$. The proof is along the lines above, however rather tedious [15] (therefore omitted), and we arrive at the following optimal filter

$$\Phi_{opt}^{a,b}(k) = \frac{v_1^{a,b} P_1^2(k) + v_2^{a,b} P_2^2(k)}{H(k)}, \qquad (8)$$

where $\vec{v}^{a,b} = (v_1^{a,b}, v_2^{a,b})^T$ are arbitrary vectors since any linear combination of diagonal matrices is diagonal.

If we set the vectors $\vec{v}^{a,b}$ as $(1,0)^T$ and $(0,1)^T$, respectively then our theory converges to the case analyzed by Pham et al. [16], since the filters reduce to $\Phi^a(k) = P_1^{-2}(k)$ and $\Phi^b(k) = P_2^{-2}(k)$.

Optimal Filtering for more than two sources The theory for optimal filtering was based on exact diagonalization of two covariance matrices. One can in principle also derive the two optimal filters $C^{a,b}$ for the N-dimensional case, resulting in more complicated implicit equations for the filters. We follow, however, a simpler heuristic approach based on the idea of pair-wise separation of sources, i.e. for each pair of source spectra $P_i(k)$ and $P_j(k)$ with $i \neq j$, we construct two filters (in general) according to (8) resulting in a total of $M(M-1)$ filters. We now diagonalize these $M(M-1)$ covariance matrices simultaneously by Jacobi-rotations, following the spirit of [7, 19]. As can be seen from (8) both in the limit of low[2] and high noise level only M of them are linear independent. In the intermediate case we perform a PCA analysis and take only the M filters corresponding to the M largest eigenvalues. The rest proceeds as before; i.e. the now M covariance matrices are approximately simultanously diagonalized.

Optimal Filtering With Noise So far we did not consider additive noise which is present for example in biomedical applications $\vec{x} = A\vec{s} + \epsilon\vec{n}$. Channel noise severely corrupts the ICA decomposition since its contributions to the covariance matrices in general does not vanish even in the limit of long averages ($T \to \infty$) [9, 14]. Under the assumption that we have some knowledge about the spectral form of $\vec{n}$, the amplitudes and all correlations of the noise, we can either subtract the noise covariance contributions or we can construct a filter under the constraint of orthogonality to the noise spectrum. For example in the case of white noise it is instructive to notice that this is equivalent to using time delayed covariance matrices only [14], since the respective frequency responses of the filters are orthogonal to the constant (flat) spectrum of white noise (cf. Fig. 1). Given that the noise contribution is properly subtracted, it is straight forward to show that the weight $H(k)$ from Eq.(8) changes as

$$\begin{aligned} H(k) \;=\; & P_1^2(k)P_2^2(k) + P_1^2(k)\overline{|N_2(k)|^2} + \\ & P_2^2(k)\overline{|N_1(k)|^2} + \frac{1}{2}\overline{\Big(G(k) - \overline{G(k)}\Big)^2}, \end{aligned} \qquad (9)$$

where $N_i(k)$ is the noise in the $i.th$ independent signal and $G(k) = N_1^*(k)N_2(k) + N_1(k)N_2^*(k)$.

If noise is dominant and white, the optimal filters read $\Phi^{a,b} = v_1^{a,b}P_1^2 + v_2^{a,b}P_2^2$, and we arrive at $\Phi^a = P_1^2$ and $\Phi^b = P_2^2$, i.e. the filters proposed in [11], however without a prior whitening step.

[2]For vanishing channel noise orthogonalization is not necessary

3. SIMULATIONS

We will now evaluate the theory described in the last section and we compare the performance of the optimal filter (OFI) algorithm to common BSS algorithms by numerical simulations.

Let us recall that the theoretical approach was based on the assumption that we were given all information about the source spectra. To make use of this method in practical source separation problems we propose the following (EM-like) iteration scheme. In a first initialization step we apply an arbitrary BSS algorithm (here TDSEP) to get a coarse separation, then in a subsequent step with the help of the estimated source spectra an estimate of the optimal filter is computed. The optimal filter is then applied to the mixtures to obtain improved estimates of the sources and their spectra. This procedure is iterated. Unless varied, the number of iterations is fixed here to three.

To estimate a power spectrum from a signal it is insufficient just to calculate the squared modulus of the Fourier transform of the data. This estimate does not converge even in the limit $T \to \infty$, and hence some kind of spectral smoothing has to be done. Here we simply apply a moving average of fixed length (31) on the estimated spectra, which we found to significantly improve the results for all considered cases. A more detailed study of the importance of smoothing techniques will be given elsewhere [15].

In order to study the behavior of the optimal filters in detail we generated a number of test data. Dataset (I) consists of snapshots (2000 samples) of Gaussian AR processes with random coefficients and variable order (between 2 and 20). Dataset (II) contains 15 real audio signals (speech, music and colored noise) sampled at 8kHz (10000 samples). All signals are artificially mixed by a randomly fixed, square matrix W. This matrix was used to compare the performance of the algorithms in the following way: Let P be a permutation matrix and D be a scaling matrix such that the unmixed vectors $\vec{u} = DPU^{-1}\vec{x}$ match the sources as good as possible. If the unmixing is successful than $E := DPU^{-1}W$ very closely resembles the identity matrix. The separation results can be measured as the deviation of E from identity by the quantity

$$p(E) = \frac{1}{M}\sqrt{\sum_{i\neq j}^{M} |E_{ij}|^2} \qquad (10)$$

where E_{ij} are the matrix elements of E.

In Fig. 2 we show the results for the optimal filter source separation algorithms (OFI Eq. (7) and OFI Eq. (8)) compared to MS [13], TDSEP [19], ARSEP

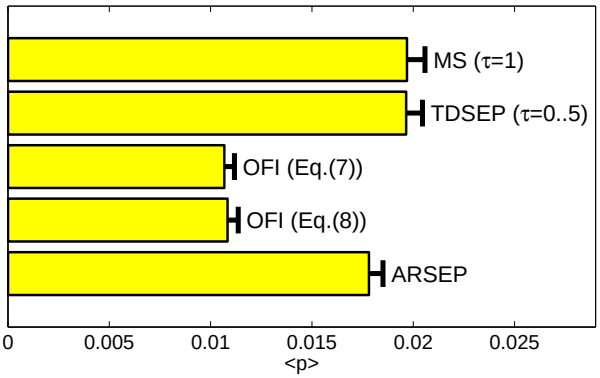

Figure 2: Mean and standard error of the mean for 300 separation trials on the AR dataset.

[11] for dataset (I) and in Fig. 3 the corresponding results for dataset (II) are presented. Here we have also included the result of the JADE algorithm [6] as a comparison to methods based on higher-order statistics.

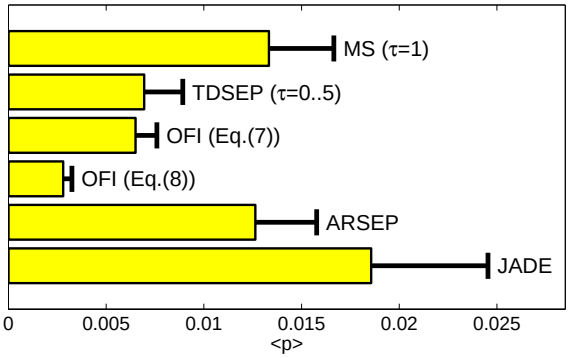

Figure 3: Mean and standard error of the mean for 60 separation trials on the audio dataset, where two signals were randomly chosen at each case.

Clearly the OFI algorithms are superior to the rest, as spectral information is used beneficially. Although the TDSEP algorithm was used to initialize the OFI algorithm the final performances are essentially independent of each other as can be seen from Fig. 4 where the performance of OFI is plotted versus the one of TDSEP. The majority of points are below the diagonal, which means superiority for OFI.

Another interesting experiment is shown in Fig. 5: the averaged performace of the iterated OFI algorithm is plotted as a function of iterations. From the curve we see first a sharp drop and then a saturation plateau.

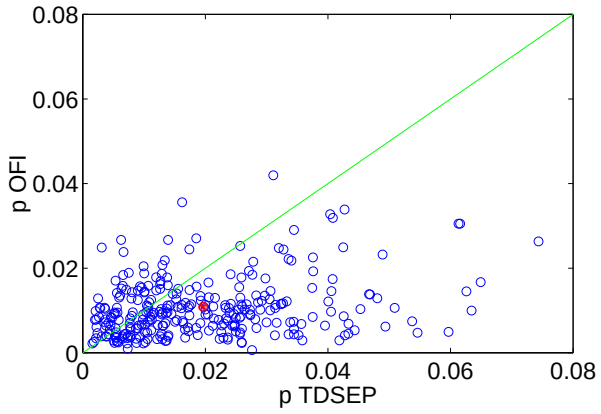

Figure 4: Scatter plot of the performance measure p for OFI and TDSEP applied to simulated AR processes. Every point corresponds to one demixing experiment.

The last simulation (cf. Fig. 6) considers the separation of more than 2 sources and we see that our approximation scheme that relies on the diagonalization of several covariance matrices, constructed from pairwise optimally filtered signals, works significantly better than the TDSEP method.

For simulations of the noise case described above we refer to [15].

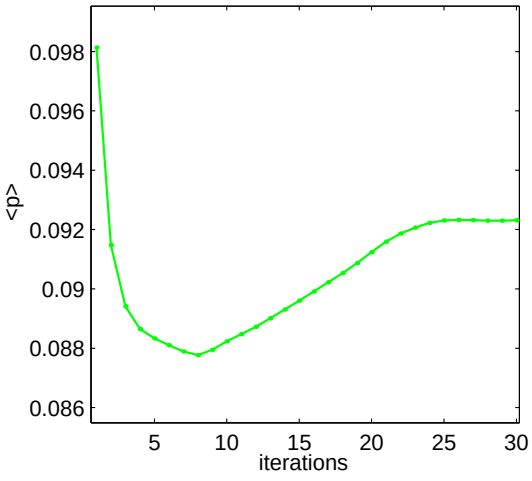

Figure 5: Average performance of OFI as a function of iterations. The average performance of the TDSEP method is 0.14.

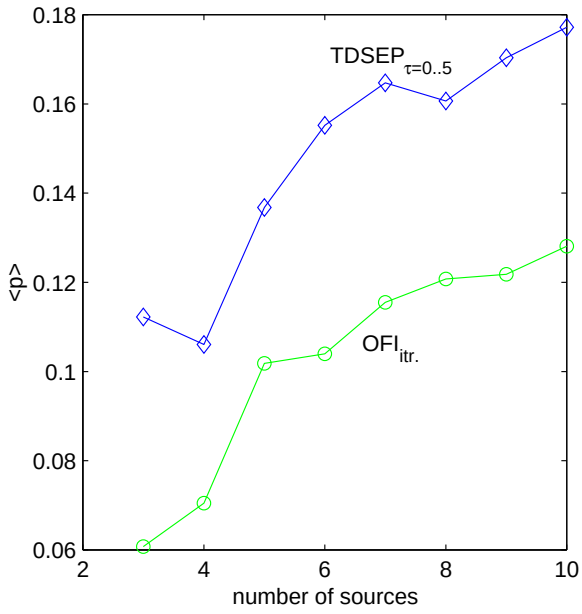

Figure 6: Average performance of TDSEP and OFI on the AR data set for increasing number of sources.

4. CONCLUSION

In this work we explained theoretically how to construct covariance matrices of *optimally* filtered signals for the best possible use of spectral information within the framework of second-order source separation algorithms. Previous second-order algorithms that make use of time-shift operations by computing delayed covariance matrices emerge as special cases of our approach and can be intuitively interpreted and implemented as particular filters. Examples are the algorithms by Tong & Liu [17], Pham & Garat [16], Belouchrani et al. [4], Molgedey & Schuster[13], Köhler & Orglmeister [11] and Ziehe & Müller [19].
We would also like to see our optimal filtering approach as a tool for benchmarking since it gives the limit that algorithms can reach by using spectral information. Furthermore our method can be seen as a way of incorporating prior knowledge about the spectra of the sources or the noise into source separation.

Numerical simulations confirmed the usefulness of our theory. The proposed iterative method for a combined estimation of filters and source signals provided remarkably good results in simulations, although the questions of stability and equivariance properties [5] deserve further study. Another aspect form the implementation point of view is that the spectra were so far

estimated empirically by taking the squared modulus of Fourier-transformed signals. However, empirically we find that with appropriate smoothing or regularization techniques one obtains better estimates and one can improve the performance even further.

Future research will therefore be dedicated to incorporate results from spectral estimation theory [8] into our OFI framework. We also plan to apply the treatment of noise as in Eq.(9) to MEG data, where parts of the system noise can be estimated in advance.

5. REFERENCES

[1] S. Amari. Estimating functions of independent component analysis for temporally correlated signals. Technical report, RIKEN BSI, May 1999.

[2] S. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In *NIPS 95*, pages 882–893. MIT Press, 1996.

[3] A. J. Bell and T. J. Sejnowski. An information maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[4] A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. on Sig. Proc.*, 45(2):434–44, February 1997.

[5] J.-F. Cardoso and S. Amari. Maximum likelihood source separation: equivariance and adaptivity. In *Proc. of SYSID'97, 11th IFAC symposium on system identification, Fukuoka, Japan*, pages 1063–1068, 1997.

[6] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140(6):362–370, 1993.

[7] J.-F. Cardoso and A. Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM J. Mat. Anal. Appl.*, 17(1):161–164, January 1996.

[8] S. Haykin. *Adaptive Filter Theory*. Prentice Hall, Englewood Cliffs, NJ, 1991.

[9] A. Hyvärinen. Noisy independent component analysis by gaussian moments. In *ICA'99*, pages 473–478, Aussois, 1999.

[10] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[11] B.-U. Köhler and R. Orglmeister. Independent component analysis using autoregressive models. In *ICA'99*, pages 359–364, Aussois, 1999.

[12] S. Makeig, T-P. Jung, D. Ghahremani, A.J. Bell, and T.J. Sejnowski. Blind separation of event-related brain responses into independent components. *Proc. Natl. Acad. Sci. USA*, 1997.

[13] L. Molgedey and H.G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Phys. Rev. Lett.*, pages 3634–3637, 1994.

[14] K.-R. Müller, P. Philips, and A. Ziehe. $JADE_{TD}$: Combining higher-order statistics and temporal information for blind source separation (with noise). In *ICA'99*, pages 87–92, Aussois, jan 1999.

[15] G. Nolte, A. Ziehe, and K.-R. Müller. Optimal filters for source separation. in preparation.

[16] D.T. Pham, P. Garat, and C. Jutten. Separation of a mixture of independent sources through a maximum likelihood approach. In J. Vandewalle, R. Boite, M. Moonen, and A. Oosterlinck, editors, *Signal Processing VI, Theories and Applications*, pages 771–774. Elsevier, 1992.

[17] L. Tong, R. Liu, and Y.H.V.C. Soon. Indeterminacy and identifiability of blind identification. *IEEE Trans. on Circuits and Systems*, 38(5):499–509, 1991.

[18] R. Vigário, V. Jousmäki, M. Hämäläinen, R. Hari, and E. Oja. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. In *NIPS'97*. MIT Press, 1998.

[19] A. Ziehe and K.-R. Müller. TDSEP – an efficient algorithm for blind separation using time structure. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *Proc. Int. Conf. on Artificial Neural Networks (ICANN'98)*, pages 675 – 680. Springer Verlag, 1998.

[20] A. Ziehe, K.-R. Müller, G. Nolte, B.-M. Mackert, and G. Curio. Artifact reduction in magnetoneurography based on time-delayed second order correlations. *IEEE Trans. Biomed. Eng.*, 47(1):75–87, 2000.