CONVOLUTIVE DECORRELATION PROCEDURES FOR BLIND SOURCE SEPARATION

R. Vollgraf[1], M. Stetter[1], and K. Obermayer[1]

[1] Dept. of Computer Science, Technical University of Berlin,
D-10587 Berlin, Germany; email: vro@cs.tu-berlin.de

ABSTRACT

Convolutive decorrelation algorithms form a class of powerful algorithms for blind source separation. In contrast to ICA, they are based on vanishing second order cross correlation functions between sources. We provide an analyze an unifying approach for convolutive decorrelation procedures. The convolutive decorrelation procedures impose the problem of simultaneously diagonalizing a number of covariance matrices. We examine different cost functions for simultaneous diagonalization with respect to the demixing matrix. It turns out, that best performance is achieved for a cost function, that takes the squared sum of the off diagonal elements after the diagonal elements were normalized to unity. We then provide criteria for convolution kernels, that are optimal for noise robustness and which can guarantee positive definite covariance matrices, which are important for reliable convergence.

1. INTRODUCTION

Blind Source Separation is an emerging signal processing technique, which aims at recovering unobserved signals or "sources" from a set of observable linear mixtures of these sources. However, no direct information about the sources or about the mixing process is available, giving rise to the adjective "blind".

A class of powerful BSS algorithms is based on the assumption of vanishing cross correlation functions between the sources. Different decorrelation algorithms belonging to that class have been published [1, 2]. Molgedey and Schuster provide in [3] a analytic solution for diagonalizing two time delayed covariance matrices $\mathbf{C_x}(\tau_1)$ and $\mathbf{C_x}(\tau_2)$, based on the evaluation of the eigensystem of $\mathbf{C_x}(\tau_1)^{-1}\mathbf{C_x}(\tau_2)$. They showed, that for ideally uncorrelated sources, two covariance matrices are sufficient to estimate the demixing matrix $\mathbf{W}$ and proposed to choose $\tau_1 = 0$, therefore this algorithm is often referred to as single shift algorithm, whereas multi shift algorithms attempt to diagonalize more than two delayed covariance matrices [4, 2]. *Dynamic Component Analysis* [5] contains decorrelation algorithms as a special case when Gaussian model distributions are used. There, the outputs of a short time Fourier-transformation are attempted to have zero cross correlation functions.

A general examination of decorrelation algorithms, however, has not yet been done. In this work, we give a unified approach to to convolutive decorrelation algorithms and provide a systematic analysis of the properties of these algorithms.

For the task of blind source separation we have to estimate an unknown mixture of also unknown sources only by observing their mixtures. Usually the mixture is supposed to be linear and noise may be added to the mixed signals.

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n} \qquad (1)$$

The bold letters $\mathbf{x}$, $\mathbf{s}$ and $\mathbf{n}$ denote signal vectors of the observations, the sources and the noise respectively. All signals are assumed to have zero mean. To estimate the mixing matrix $\mathbf{A}$ or its inverse $\mathbf{W}$, the sources must have certain properties, which allow them to be detected in the mixture. In ICA algorithms this is the assumption of independently distributed source amplitudes [6].

Decorrelation procedures make use of temporal[1] correlations up to moments of second order. For uncorrelated sources, which is the model assumption for this class of source separating algorithms, temporal correlations occur in each source signal but not among different sources.

$$\langle s_i(t)s_k(t+\tau)\rangle = 0 \qquad (2)$$

for $i \neq k$ and any τ. For the correlations between sources we introduce the matrix notation $\mathbf{C_s}(\tau) = \langle \mathbf{s}(t)\mathbf{s}^T(t+\tau)\rangle$. We often refer to $\mathbf{C_s}(\tau)$ as *delayed covariance matrix*, in contrast to the covariance matrix in the statistical sense, which would be $\mathbf{C_s}(0)$. As well, we refer to signals as uncorrelated, if the cross correlations vanish for all τ.

[1] or spatial, for signals in a spatial domain

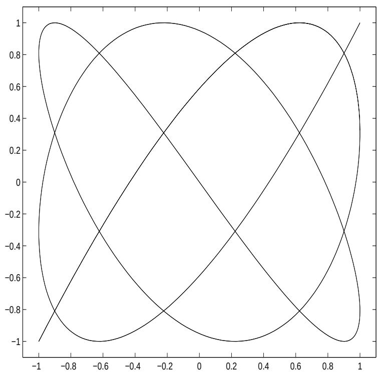

Figure 1: Amplitudes of two cosine signals with a frequency ratio of 7:5. The signals have zero cross correlation functions but are not independent.

The source property for decorrelation algorithms is therefore diagonal covariance matrices $\mathbf{C_s}(\tau)$ for any τ. To obtain non zero diagonal elements of $\mathbf{C_s}(\tau)$ for $\tau \neq 0$, the sources must not have delta peaked autocorrelation functions, which imposes smoothness in the sources. Also, all sources must not have identical autocorrelation functions, which would lead to linearly dependent $\mathbf{C_s}(\tau)$ for all τ. This model is not equivalent to the model assumption of ICA. As shown in figure 1, it is easy to construct two signals which have zero cross correlation functions but are not independent. Similarly independent signals do not need to have zero cross correlation functions, since no temporal information is involved to the model of ICA. Consider a signal with temporarily i.i.d. amplitudes. This signal and it's by $\tau \neq 0$ delayed version are independent but have correlations at τ.

2. GENERAL APPROACH

In this section we present a unifying approach to decorrelation based blind source separation. For source separation we want to find a demixing matrix $\mathbf{W}$, which yields signals $\mathbf{y}$ with vanishing cross correlation functions.

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{s} \tag{3}$$

Because of the linearity of the transformations, it holds

$$\mathbf{C_y}(\tau) = \mathbf{W}\mathbf{C_x}(\tau)\mathbf{W}^T = \mathbf{W}\mathbf{A}\,\mathbf{C_s}(\tau)\,(\mathbf{W}\mathbf{A})^T, \tag{4}$$

where $\mathbf{C_x}(\tau)$ and $\mathbf{C_y}(\tau)$ denote the shifted covariance matrices of the mixtures and the recovered sources respectively. Apart from trivial solutions, $\mathbf{C_y}(\tau)$ is only diagonal for arbitrary τ when $\mathbf{W}\mathbf{A}$ is a permutation and scaling matrix. From there, the task of source separation is as following:

- Observe a number of delayed covariance matrices $\mathbf{C_x}(\tau_l)$ at different τ_l.
- Adjust $\mathbf{W}$ such that all $\mathbf{C_y}(\tau_l)$ become diagonal.

As we shall see shortly, signals of finite length, in general, have only approximately vanishing cross correlation functions. Therefore, the choice of the shift parameters τ is crucial for the success of the algorithm. Optimal τ would yield minimal cross correlations compared to the autocorrelations of the sources. Since the sources and hence optimal τ are unknown, one can achieve a kind of averaging by using more then two shifts, which leads to *multi shift algorithms* [4].

A more general approach however, is to consider filtered or convolved versions of the signals instead of time delayed ones. For time discrete signals, the convolution of a signal x_i with a kernel f is given by

$$(x_i * f)(t) = \sum_{\tau} x_i(t-\tau) f(\tau). \tag{5}$$

It is easy to see, that with $f(\tau) = \delta(\tau - \tau_0)$, a shift is a special case of a convolution. Our unified approach to blind source separation based on decorrelation uses correlation matrices of a set of convolved versions of mixtures for diagonalization. Therefore, we introduce the *convolved covariance matrix*

$$\mathbf{C}_{\mathbf{x}}^{f,g} = \left\langle (\mathbf{x} * f)(\mathbf{x} * g)^T \right\rangle, \tag{6}$$

where $(\mathbf{x} * f)$ denotes the component wise convolution of the signal vector $\mathbf{x}$ with f. For source separation, a demixing matrix $\mathbf{W}$ has to be found that solves the following set of equations

$$\mathbf{C}_{\mathbf{y}}^{f,g} = \mathbf{W}\mathbf{C}_{\mathbf{x}}^{f,g}\mathbf{W}^T = \mathbf{\Lambda}^{(f,g)} \tag{7}$$

for all considered pairs of convolution kernels (f,g). $\mathbf{\Lambda}^{(f,g)}$ is diagonal matrix, depending on f and g. The convolved covariance matrix $\mathbf{C}_{\mathbf{y}}^{f,g}$ is a linear combination of all time delayed covariance matrices, weighted with the values of the convolution kernels f and g at the corresponding delays

$$\mathbf{C}_{\mathbf{y}}^{f,g} = \sum_{\tau_1}\sum_{\tau_2} f(\tau_1) g(\tau_2) \mathbf{C_y}(\tau_2 - \tau_2). \tag{8}$$

$\mathbf{C}_{\mathbf{y}}^{f,g}$ can only be diagonal, independently from f and g, if $\mathbf{C_y}(\tau)$ is diagonal for all τ. Hence, it follows, that **signals have zero cross correlation functions if, and only if, arbitrary convolved or filtered versions of the signals are uncorrelated at zero time**

delay. The single shift (and the zero shift) used in [3] are special cases of convolutions with δ-functions. For Dynamic Component Analysis [5], a short time Fourier-transformation can be considered as a convolution with a finite piece of a sinusoidal signal. Thus, this algorithm can be seen as a convolutive decorrelation algorithm using symmetric pairs (f, f) of convolution kernels. In the following sections, we analyse some properties of the presented unified approach.

2.1. Problems with short signals

The expectation in (2) is known as the cross correlation function between signal s_i and s_k. Both, the signals and their cross correlation function may be transformed into the frequency domain using Fourier-Transformation. In this work, we denote quantities in the frequency domain with the corresponding upper case letters. For infinite, time discrete signals, the Fourier-Transformation is

$$S(j\omega) = \sum_t s(t) e^{(-j\omega t)}, \qquad (9)$$

where j is the imaginary unit. The transformation to the frequency domain allows the cross correlation function to be easily computed by a conjugate complex multiplication. So, the property of uncorrelated sources (2) is in the frequency domain

$$S_i(j\omega) S_k(j\omega)^C = 0 \qquad \text{for} \quad i \neq k, \qquad \text{and any} \quad \omega, \qquad (10)$$

where $(\cdot)^C$ denotes the conjugate complex of the argument. That is, the frequency components of uncorrelated signals are sparse, i.e. for any ω no more than one signal may be active.

A signal of finite length can be considered as a infinite signal, multiplied with a window function, which is different from zero only for a range of finite length. The multiplication in the time domain becomes a convolution in the frequency domain. Hence, the spectra of the signals get blurred with the spectrum of the window function. Hence, (10) holds only approximately and the cross correlation functions dont vanish completely.

Another problem arises from the fact, that the covariance matrices $\mathbf{C_s}(\tau)$ and $\mathbf{C}_{\mathbf{s}}^{f,g}$ of finite signals are not symmetric by construction. So, if they are not diagonal, they are not necessarily symmetric. Therefore, the covariance matrices of any linearly transformed signals might be not symmetric either. This turns out to be a problem for a number of algorithms, in particular for those making use of eigenvalues, which can be complex for not symmetric matrices. Any matrix $\mathbf{C}$ can be split however, into a sum of a symmetric and a antisymmetric matrix, using $\mathbf{C}^{(sym)} = \frac{1}{2}(\mathbf{C} + \mathbf{C}^T)$ and $\mathbf{C}^{(asym)} = \frac{1}{2}(\mathbf{C} - \mathbf{C}^T)$. To diagonalize $\mathbf{C}$ means to diagonalize the symmetric part and make the antisymmetric part vanish. However, it turns out, that for a linear transformation with real $\mathbf{W}$, the antisymmetric part $\left(\mathbf{W}\mathbf{C}^{(asym)}\mathbf{W}^T\right) = -\mathbf{W}\mathbf{C}^{(asym)}\mathbf{W}^T$ stays antisymmetric and cannot vanish for nonsingular $\mathbf{W}$. Because source estimates must be real, the antisymmetric part of the covariance matrices can be regarded as a finite-size effect of the signals. It must vanish in the limit of infinite signal length. Thus, we drop the antisymmetric part of $\mathbf{C}$ and **from now on, consider all covariance matrices as symmetric.**

3. COST FUNCTIONS

In this section, we investigate, which properties of cost functions are advantageous for the efficient simultaneous diagonalization of the equations (7). For diagonalization of more than two covariance matrices, an analytic solution like in [3] does not exist anymore. Further, since the sources are usually only approximately uncorrelated, the best achievable solution would still not completely diagonalize the covariance matrices. Therefore, a measure is needed, which evaluates, "how strongly diagonal" the covariance matrices of the reconstructed signals $\mathbf{C}_{\mathbf{y}}^{f,g} = \mathbf{W}\mathbf{C}_{\mathbf{x}}^{f,g}\mathbf{W}^T$ are. This measure is implemented by a cost function, which has to be minimized with respect to $\mathbf{W}$.

3.1. Simple cost function

One simple choice for a cost function penalizes the squared sum over all off-diagonal elements of $\mathbf{C}_{\mathbf{y}}^{f,g}$.

$$E = \sum_{\{(f,g)\}} \sum_i \sum_{j \neq i} \left(\mathbf{W}\mathbf{C}_{\mathbf{x}}^{f,g}\mathbf{W}^T\right)_{ij}^2, \qquad (11)$$

where the first sum runs over all pairs off convolution kernels (f, g). The gradient off (11) has the form

$$\nabla E = 4 \sum_{\{(f,g)\}} \left(\mathbf{C}_{\mathbf{y}}^{f,g} - diag(\mathbf{C}_{\mathbf{y}}^{f,g})\right) \mathbf{W}\mathbf{C}_{\mathbf{x}}^{f,g}. \qquad (12)$$

It can be easily seen, that the gradient vanishes if all $\mathbf{C}_{\mathbf{y}}^{f,g}$ are diagonal. Unfortunately, the gradient may also vanish, when $\mathbf{W}$ decays completely to zero, which would be the only fix point of the cost function in the case of not ideally uncorrelated sources.

Hence, it is necessary to either constrain $\mathbf{W}$ explicitly, to prevent it from growing to small, or to setup a cost function, which is invariant on changes of the variance of the outputs $\mathbf{y}$.

3.2. Constraining the diagonal of $\mathbf{W}^{-1}$

Molgedey and Schuster propose in [3] also a cost function for the use with gradient based optimization. They compare their algorithm with a recurrent network of linear neurons, which have inhibitory connections between each other, but not among themselves. The matrix $\mathbf{T}$ contains these connection and has a zero diagonal. The equivalent feed-forward network of this architecture is $\mathbf{W} = (\mathbf{1} + \mathbf{T})^{-1}$. From the view of the mixing process, this means that to each source, a linear combination of all other sources is added. This leads to the constraint

$$\mathbf{W}_{ii}^{-1} =_{def.} 1, \tag{13}$$

and the cost function (11) is minimized with respect to the elements of $\mathbf{T}$. However, this constraint does not prevent $\mathbf{W}$ from getting arbitrary small. Consider a matrix

$$\mathbf{W}^{-1} = \begin{pmatrix} 1 & a & \cdots & a \\ a & 1 & \cdots & a \\ \vdots & \vdots & \ddots & \vdots \\ a & a & \cdots & 1 \end{pmatrix},$$

for which (13) is fulfilled. For large a, the corresponding matrix $\mathbf{W}$ is

$$\mathbf{W} = \frac{1}{a} \begin{pmatrix} -\frac{N-2}{N-1} & \frac{1}{N-1} & \cdots & \frac{1}{N-1} \\ \frac{1}{N-1} & -\frac{N-2}{N-1} & \cdots & \frac{1}{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{N-1} & \frac{1}{N-1} & \cdots & -\frac{N-2}{N-1} \end{pmatrix},$$

where N is the number of sources. Increasing the value of a will decrease the row norm of $\mathbf{W}$ and hence decrease the cost function without diagonalizing $\mathbf{C}_{\mathbf{y}}^{f,g}$. The following example shall illustrate the presence of trivial minima. Consider two covariance matrices $\mathbf{C}_{\mathbf{s}}^{1} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ and $\mathbf{C}_{\mathbf{s}}^{2} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ belonging to two source signals, which are mixed with $\mathbf{A} = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$. Figure 2 shows a contour plot of the cost function with respect to the two off-diagonal elements of $\mathbf{W}^{-1}$. The hyperbolic ridge corresponds to values at which $\mathbf{W}^{-1}$ becomes singular. Solid lines mark gradient descent trajectories for 5 different initializations. It can be seen, that depending on the initialization, the gradient descent procedure may succeed (trajectories 1,3) or fail to converge to the wanted minimum, but instead diverge, leading to arbitrary small $\mathbf{W}$ and hence trivial minima (trajectories 2,4,5). Therefore, we consider the constraint (13) as not suitable for gradient based joint diagonalization.

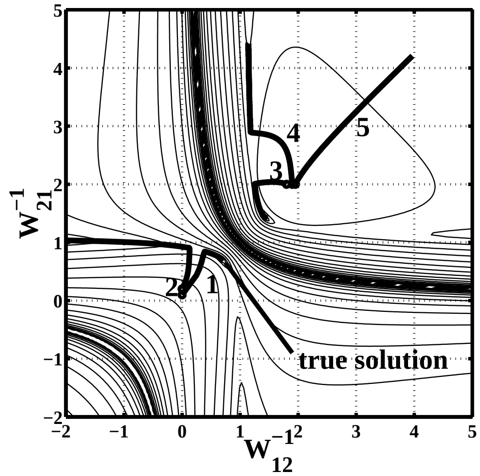

Figure 2: Surface of the cost function with the constraint $\mathbf{W}_{ii}^{-1} = 1$ for two sources. The two hyperbolic ridges are regions with singular $\mathbf{W}^{-1}$ and have infinite values of the cost function.

3.3. Normalized covariance matrices

The spurious minima of the simple cost function (11) arise from vanishing variances of the output signals $\mathbf{y}$. To avoid this, we examined a modification of the cost function, which normalizes the off-diagonal elements of $\mathbf{C}_{\mathbf{y}}^{f,g}$ by the diagonal elements. Thus, the gradient of the cost function is zero for directions, which only influence the row norm of $\mathbf{W}$ and hence the output variances. The modified cost function is given by

$$E = \sum_{\{(f,g)\}} \sum_{i} \sum_{j \neq i} \frac{\left(\mathbf{C}_{\mathbf{y}}^{f,g}\right)_{ij}^{2}}{\left|\left(\mathbf{C}_{\mathbf{y}}^{f,g}\right)_{ii} \left(\mathbf{C}_{\mathbf{y}}^{f,g}\right)_{jj}\right|}. \tag{14}$$

Since the $\mathbf{C}_{\mathbf{y}}^{f,g}$ are not necessarily positive definite, the normalizing term must appear as absolute value in the denominator of (14). The gradient can be obtained from

$$\frac{\partial E}{\partial \mathbf{W}_{kl}} = 2 \sum_{\{(f,g)\}} \sum_{j \neq k} \frac{\mathbf{C}_{kj}}{\left|\mathbf{C}_{kk}\mathbf{C}_{jj}\right|} \left(\left(\mathbf{W}\mathbf{C}_{\mathbf{x}}^{f,g}\right)_{jl} - 2\frac{\mathbf{C}_{kj}}{\mathbf{C}_{kk}} \left(\mathbf{W}\mathbf{C}_{\mathbf{x}}^{f,g}\right)_{kl} \right). \tag{15}$$

$\mathbf{C}_{kj}$ abbreviates $\left(\mathbf{C}_{\mathbf{y}}^{f,g}\right)_{kj}$. Simulations have shown, that this cost function performs well as long as all covariance matrices are positive definite. In two experiments, 80 simultaneous diagonalizations of four 4×4 covariance matrices each, where performed, using positive definite $\mathbf{C}_{\mathbf{s}}^{f,g}$ in one experiment and not positive definite $\mathbf{C}_{\mathbf{s}}^{f,g}$ in the other. The four covariance matrices where initialized with uniformly distributed diagonal elements $\left(\mathbf{C}_{\mathbf{s}}^{f,g}\right)_{ii} \in (0 \cdots 1)$ rsp. $\left(\mathbf{C}_{\mathbf{s}}^{f,g}\right)_{ii} \in (-1 \cdots 1)$ and mixed with $\mathbf{A}$, which was initialized with elements

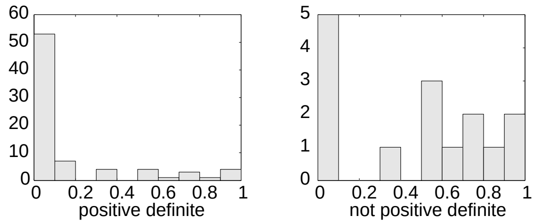

Figure 3: Histogram of the reconstruction error of the successful trials (finite reconstruction error) with pos. def. $\mathbf{C}_{\mathbf{s}}^{f,g}$ (left) and not pos. def. $\mathbf{C}_{\mathbf{s}}^{f,g}$ (right). Positive definite covariance matrices perform considerably better.

from a (0,1)-normal distribution. The results were evaluated, using the reconstruction error. This quantity gives a measure of the ratio of the off-permutation elements of $\mathbf{WA}$ to the permutation elements. The permutation is determined by the largest elements in each row. If $\mathbf{WA}$ cannot be interpreted as a permutation matrix, the source separation is considered as failed and the reconstruction error assigned the value infinity.

In the experiment with the positive definite $\mathbf{C}_{\mathbf{s}}^{f,g}$, a finite reconstruction error could be achieved in 77 of 80 trials, whereas with the not positive definite $\mathbf{C}_{\mathbf{s}}^{f,g}$ only 15 of 80 separations where successful.

Figure 3 shows the histogram of the reconstruction error of the successful trials. It is on average much lower if all matrices $\mathbf{C}_{\mathbf{s}}^{f,g}$ are positive definite.

3.4. Cost function based eigenvalues

Another cost function that is independent from the output variances can be found, if the diagonal elements of $\mathbf{C}_{\mathbf{y}}^{f,g}$ are compared with their eigenvalues.

$$E = \sum_{\{(f,g)\}} \mathcal{D}\left(diag\left(\mathbf{C}_{\mathbf{y}}^{f,g}\right), eig\left(\mathbf{C}_{\mathbf{y}}^{f,g}\right)\right) \qquad (16)$$

$\mathcal{D}$ is a distance measure between the vector of the eigenvalues and the vector of diagonal elements of $\mathbf{C}_{\mathbf{y}}^{f,g}$. For diagonal $\mathbf{C}_{\mathbf{y}}^{f,g}$ these two vectors are, if properly assigned, identical. To obtain invariance on output variances, the diagonal elements need to be normalized to unity

$$\left(\tilde{\mathbf{C}}_{\mathbf{y}}^{f,g}\right)_{ij} = \left(\mathbf{C}_{\mathbf{y}}^{f,g}\right)_{ij} \left|\left(\mathbf{C}_{\mathbf{y}}^{f,g}\right)_{ii}\left(\mathbf{C}_{\mathbf{y}}^{f,g}\right)_{jj}\right|^{-\frac{1}{2}}. \qquad (17)$$

Thus, the cost function has the form

$$E = \sum_{\{(f,g)\}} \left(\sum_i \left|1 - |\lambda_i^{f,g}|\right|^{\alpha_1}\right)^{\alpha_2}, \qquad (18)$$

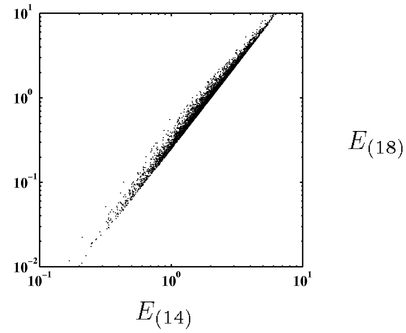

Figure 4: Logarithmic plot of the values of the cost functions $E_{(14)}$ and $E_{(18)}$ for 2000 points near the global minimum.

where the parameters α_1 and α_2 control the distance metric and $\lambda_i^{f,g}$ are the eigenvalues of $\tilde{\mathbf{C}}_{\mathbf{y}}^{f,g}$.

This cost function performs quite similar to (14) as shown in figure 4, where for 2000 separating matrices near the absolute minimum $\mathbf{A}^{-1}$ of the experiment of figure 2, the values of both cost functions are plotted against each other. The narrow shape of the cloud corresponds to a strong correlation between the two cost functions. However (18) is quite computationally expensive, due to the need of numerically calculating the gradient, which includes multiple evaluations the eigensystems of $\mathbf{C}_{\mathbf{y}}^{f,g}$.

In summary, we recommend the use of cost function (14), because it is fast to compute and for positive $\mathbf{C}_{\mathbf{y}}^{f,g}$ avoids spurious minima corresponding to trivial solutions.

4. CONVOLUTION KERNELS

For convolutive decorrelation algorithms we may freely choose the set set of convolutions f and g, which we want to use for source separation. In this section we provide two criteria for good choices of convolution kernels.

4.1. Noise robustness

In the presence of white noise, a noise vector is added to the linear mixture $\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n}$. The noise signals are considered to have zero cross correlation functions to each other and all other signals, i.e. $(\mathbf{C}_{\mathbf{n}}(\tau))_{ik} = 0$ for all τ and all $i \neq k$. Thus, the best source separation result for $\mathbf{W} = \mathbf{A}^{-1}$ yields $\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{s} + \mathbf{A}^{-1}\mathbf{n}$. That is, in the best case, the sources are properly separated, but the noise is still present. In the general case, the covariance matrices of the outputs can be written as

$$\mathbf{C}_{\mathbf{y}}^{f,g} = \mathbf{C}_{\mathbf{s}}^{f,g} + \mathbf{W}\mathbf{C}_{\mathbf{n}}^{f,g}\mathbf{W}^T. \qquad (19)$$

Optimal convolution kernels for noise robustness have to suppress $\mathbf{C}_{\mathbf{n}}^{f,g}$, which is, equivalently to (8),

$$\mathbf{C}_{\mathbf{n}}^{f,g} = \sum_{t_1}\sum_{t_2} f(-t_1)g(-t_2)\mathbf{C}_{\mathbf{n}}(t_2 - t_1) = 0 \qquad (20)$$

The noise is assumed to be white, thus $\mathbf{C}_{\mathbf{n}}(t_2 - t_1) = 0$ for $t_1 \neq t_2$. One can see, that the noise term in (19) is zero when f and g are orthogonal.

$$\sum_t f(t)g(t) = 0 \qquad (21)$$

4.2. Positive definite $\mathbf{C}_{\mathbf{y}}^{f,g}$

For the use of cost function (14), it is important to have positive definite covariance matrices. In this section, we examine if one can generally find convolution kernels, that guarantee positive definite $\mathbf{C}_{\mathbf{y}}^{f,g}$ and $\mathbf{C}_{\mathbf{s}}^{f,g}$. Because only linear transformations are performed, for positive definite $\mathbf{C}_{\mathbf{y}}^{f,g}$ can be checked at the observable $\mathbf{C}_{\mathbf{x}}^{f,g}$. A guarantee would give a symmetric pair of convolution kernels $f = g$. This, however, can never be orthogonal and thus, is not able to suppress the noise term. The goal is to achieve positive diagonal elements of $\mathbf{C}_{\mathbf{s}}^{f,g}$. The i-th diagonal element is the correlation of the two signals $s_i * f$ and $s_i * g$. This quantity can be calculated in the frequency domain

$$\left(\mathbf{C}_{\mathbf{s}}^{f,g}\right)_{ii} = \frac{1}{2\pi}\int_{-\infty}^{+\infty} F(j\omega)G(j\omega)^C \, S_i(j\omega)S_i(j\omega)^C \, d\omega \qquad (22)$$

Making use of the conjugate complex symmetry of the spectrum of real signals, we obtain the desired condition

$$\Re\left(F(j\omega)G(j\omega)^C\right) \geq 0 \qquad \text{for all} \quad \omega. \qquad (23)$$

The equality must not hold for at least one ω with $S_i(j\omega) \neq 0$, otherwise we would get a zero diagonal element in $\mathbf{C}_{\mathbf{s}}^{f,g}$ (and also in $\mathbf{C}_{\mathbf{y}}^{f,g}$) which would lead to a singularity in the normalization term of cost function (14). Unfortunately, this is incompatible with (21), what can be seen from

$$\sum_t f(t)g(t) = \left.(F * G)\right|_{\omega=0} = \sum_k \Re\left(F(j\omega_k)G(j\omega_k)^C\right). \qquad (24)$$

(24) can only be zero if all sum terms on the right hand side are zero or negative terms are allowed. Thus, positive definite covariance matrices cannot be generally guaranteed with orthogonal (f, g). However, this does not mean, that it is impossible to obtain positive definite covariance matrices from orthogonal convolution kernels.

5. SUMMARY AND CONCLUSIONS

We formulated a unifying approach to blind source separation by second order decorrelation algorithms. It has been shown, that the simultaneous diagonalization of time delayed covariance matrices is a special case of the more general case of convolutive covariance matrices. The use of convolution instead of shifts can provide an averaging for approximately uncorrelated sources. We showed, that signals of finite length have in general only approximately vanishing cross correlation functions, the signal length is crucial for the model assumption. For white source signals, i.e. for delta peaked auto correlation function, the algorithms must fail. Also for all source autocorrelation functions being identical, the algorithm cannot work. As one can see from (8), all $\mathbf{C}_{\mathbf{s}}^{f,g}$ would be linearly dependent, no more than a PCA could be performed. Different cost functions were examined, and it turned out, that cost function (14) performed best. A necessary condition for reliable convergence of (14) is positive definite covariance matrices $\mathbf{C}_{\mathbf{y}}^{f,g}$. Without knowledge of the sources, (24) can guarantee positive definiteness, is however incompatible with orthogonal convolution pairs (21), which is necessary for noise robustness. If prior knowledge about the sources is available (e.g. if they are restricted to a certain frequency band), this knowledge can be used together with (22 for the construction of convolution kernels, which fulfill (21) and (24) simultaneously.

6. ACKNOWLEDGMENT

This work has been supported by German Science Foundation (grants: DFG Ob 102/3-1, DFG Se 931/1-1)

7. REFERENCES

[1] J. C. Platt and F. Faggin, "Networks for separation of sources that are superimposed and delayed," in *Advances in Neural Information Processing Systems*, S. J. Hanson J. E. Moody and R. P. Lippmann, Eds., 1991, vol. 4, pp. 730–737.

[2] M. Stetter, I. Schießl, T. Otto, F. Sengpiel, M. Hübener, T. Bonhoeffer, and K. Obermayer, "Principal component analysis and blind separation of sources for optical imaging of intrinsic signals.," *NeuroImage*, p. in press., 2000.

[3] L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," *Phys. Rev. Lett.*, vol. 72, pp. 3634–3637, 1994.

[4] H. Schöner, M. Stetter, I. Schießl, J. E. W. Mayhew, J. S. Lund, N. McLoughlin, and K. Obermayer, "Networks for separation of sources that are superimposed and delayed," in *Application of blind separation of sources to optical recording of brain activity*, T. K. Leen S. A. Solla and K.-R. Müller, Eds. 2000, pp. 949–955, MIT Press.

[5] H. Attias and C. E. Schreiner, "Blind source separation and deconvolution: The dynamic component analysis algorithm," *Neural Comput.*, vol. 10, pp. 1373–1424, 1998.

[6] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution.," *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.