Blind separation of sources that have spatiotemporal variance dependencies

Aapo Hyvärinen [a,b,*] Jarmo Hurri [a]

[a]*Neural Networks Research Centre, Helsinki University of Technology, Finland*

[b]*Helsinki Institute for Information Technology, Basic Research Unit, Dept of Computer Science, University of Helsinki, Finland*

Abstract

In blind source separation methods, the sources are typically assumed to be independent. Some methods are also able to separate dependent sources by estimating or assuming a parametric model for their dependencies. Here, we propose a method that separates dependent sources without a parametric model of their dependency structure. This is possible by introducing some general assumptions on the structure of the dependencies: the sources are dependent only through their variances (general activity levels), and the variances of the sources have temporal correlations. The method can be called double-blind because of this additional blind aspect: We do not need to estimate (or assume) a parametric model of the dependencies, which is in stark contrast to most previous methods.

Key words: independent component analysis, blind source separation, dependent component analysis, higher-order cumulants

1 Introduction

Blind source separation is typically based on the assumption that the observed signals are linear superpositions of underlying hidden source signals. Let us denote the n source signals by $s_1(t),\ldots,s_n(t)$, and the observed signals by $x_1(t),\ldots,x_m(t)$. Let a_{ij} denote the coefficients in the linear mixing between the source $s_j(t)$ and the observed signal $x_i(t)$. Further, let us collect the source signals in a vector $\mathbf{s}(t) = (s_1(t),\ldots,s_n(t))^T$, and similarly we construct the observed signal vector $\mathbf{x}(t)$. Now

* Helsinki Institute for Information Technology / BRU, P.O.Box 26, FIN-00014 University of Helsinki, Fax: +358-9-191 4441, email: aapo.hyvarinen@helsinki.fi

URL: `http://www.cs.helsinki.fi/aapo.hyvarinen/` (Aapo Hyvärinen).

the mixing can be expressed as the equation

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \tag{1}$$

where the matrix $\mathbf{A} = [a_{ij}]$ collects the mixing coefficients. No particular assumptions on the mixing coefficients are made. Some weak structural assumptions are often made, however: for example, it is typically assumed that the mixing matrix is square, that is, the number of source signals equals the number of observed signals ($n = m$), which we will assume here as well.

The problem of blind source separation is now to estimate both the source signals $s_i(t)$ and the mixing matrix $\mathbf{A}$, based on observations of the $x_i(t)$ alone [16]. The word "blind" refers primarily to the impossibility of directly observing the source signals. If the source signals could be partly observed (during some limited teaching period, for example), the problem could be solved by basic linear regression techniques. However, more sophisticated unsupervised methods are needed here; they are based on somewhat unconventional statistical properties of the source signals as will be discussed next.

In most methods, the source signals are assumed statistically independent. Then, the model can be estimated if the source signals fulfill some additional assumptions, two of which are commonly used. First, if all the components (except perhaps one) have nongaussian distributions, the ensuing model is called independent component analysis [4], and many techniques are available for estimation of the model [14]. Second, if the components have nonstationary, smoothly changing variances [17,19,10], the model can be estimated as well. (See Discussion for further possibilities.)

Recently, several researchers have considered the case where the source signals are *not* independent. Many different variants can be considered: the components might be divided into groups so that components inside a group are dependent but components in different groups are independent [3,11], the dependencies might follow topographic organization [12], the structure of trees [1], or some general parametric forms [8,21]. The dependencies either need to be exactly known beforehand, or they can be estimated as part of the method as in [1,8,21]. Each model extends the blind source separation ability to situations in which the source signals follow the prescribed parametric dependency structure (an exception being [11] where actual separation is not possible).

What we propose in this paper is a method that separates dependent sources *without a parametric model of their dependency structure*. The main assumptions are that the sources are dependent only through their variances (general activity levels), and that the variances of the sources have temporal correlation; this is what we call spatiotemporal variance dependencies. The method can be called *double-blind* in the sense that we neither observe the source signals, nor need estimate (or postulate) a parametric model of their dependencies. Certainly, assumptions on the general structure of the dependencies must be made — just as in the basic case of ICA, where the sources must be assumed nongaussian and independent.

First, we motivate and define the general kind of dependencies allowed for the source signals (Section 2). Then we propose a cumulant-based criterion, and prove that it separates the signals (Section 3). Section 4 shows simulation results, and Section 5 discusses connections to other methods and concludes the paper.

2 Model with spatiotemporal variance dependencies

2.1 Motivation

Many signals have a smoothly changing, nonstationary variance [17]. For example, if a signal is characterized by long periods of silence interspersed with bursts of activity, one can consider the signal as having a variance signal that is (close to) zero most of the time, the bursts corresponding to nonzero values. Such a behaviour is clearly seen in speech signals [17] and natural video signals [9], for example.

In models of nonstationary variance, it is conventional to model a source signal $s_i(t)$ as a product of an underlying i.i.d. signal $y_i(t)$ and a smoothly changing variance signal $v_i(t)$ [19]. Thus, we define

$$s_i(t) = v_i(t)y_i(t). \tag{2}$$

On the other hand, the variances $v_i(t)$ are often dependent among different signals, as has been observed in natural images [23] and magnetoencephalographic data [22,12], for example. This leads to a specific form of dependencies, and could be modelled by considering that the variance signals themselves are the results of a mixing process [12,21].

Thus, combination of these two properties leads to what can be called spatiotemporal variance dependencies. Previously, we pointed out that such dependencies exist in natural image sequences [8,13]. A simple artificial example of signals with such dependencies is shown in Fig. 1.

2.2 Definition of model

Based on the above motivations, we define the following signal model. The observed signals are linear mixtures of source signals as in Eq. (1) with a square mixing matrix. As usual, we assumed that the signals $s_i(t)$ have zero mean and unit

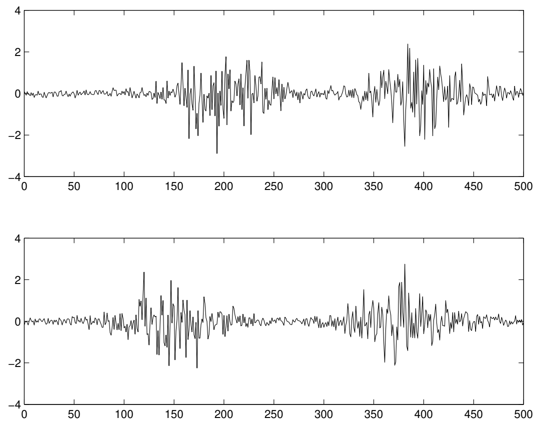

Fig. 1. A caricature of two signals that spatiotemporal variance dependencies. The variances, i.e. activity levels of the signals have temporal correlations, and also correlations between signals.

variance. Further, we assume that the sources $s_i(t)$ have dependencies because the general activity levels, i.e. variances of the sources are not independent. Moreover, we assume that these activity levels change smoothly in time. To model such dependencies, we assume that each source signal can be represented as a product of two random signals $v_i(t)$ and $y_i(t)$ as in Eq. (2). Thus, we obtain for each observed signal $x_i, i = 1 \dots n$:

$$x_i(t) = \sum_{j=1}^{n} a_{ij} v_j(t) y_j(t). \tag{3}$$

Here, $y_i(t)$ is an i.i.d. signal that is completely independent in time, and different y_i's are mutually independent over the index i as well. No assumption on the distribution of $y_i(t)$ is made, other than it must have zero mean. The signals $y_i(t)$ are also independent of the signals $v_i(t)$.

The dependencies, both between the sources and over time, are thus only due to the dependencies between (and in) the $v_i(t)$, which are nonnegative signals giving the general activity levels. Thus, $v_i(t)$ and $v_j(t)$ are allowed to be statistically dependent. No particular assumptions on these dependencies are made, in order to have as blind a method as possible. (But a condition of full rank is necessary as will be seen below.)

Our method uses the time structure of the signals, so the source signals are assumed to have some time dependencies, i.e. the signals $v_i(t)$ must have some kind of autocorrelations. An exact condition will be given in the next section.

3 A contrast function for the model

In this section, we propose a simple cumulant-based objective function whose maximization is shown to enable the estimation of the model.

We assume that the data is preprocessed by temporal and spatial whitening. First, each of the observed signals $x_i(t)$ is temporally filtered [1] by a filter that makes $x_i(t)$ and $x_i(t')$ uncorrelated for any $t \neq t'$. Then, ordinary spatial whitening, which is a standard preprocessing technique in ICA [14], is applied. The preprocessed signals are denoted by $z_i(t)$.

The contrast function is given by the following Theorem, proven in Appendix A:

Theorem 1 *Assume that the signals $x_i(t)$ are generated as described in Eq. (3), and that the signals are preprocessed by spatial whitening to give the multidimensional signal $\mathbf{z}(t)$. Define the objective function:*

$$J(\mathbf{W}) = \sum_{i,j} \left[\operatorname{cov}([\mathbf{w}_i^T \mathbf{z}(t)]^2, [\mathbf{w}_j^T \mathbf{z}(t-\Delta t)]^2)\right]^2 \tag{4}$$

where $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)^T$ is constrained to be orthogonal, and the lag Δt is nonzero. Assume that the matrix $\mathbf{K}$ defined as

$$\mathbf{K}_{ij} = \operatorname{cov}(s_i^2(t), s_j^2(t-\Delta t)) \tag{5}$$

has full rank. Then, the objective function J is (globally) maximized when $\mathbf{WA}$ equals a signed permutation matrix, i.e. the $\mathbf{w}_i^T \mathbf{z}(t)$ equal the original sources $s_i(t)$ up to random signs.

This is a generalization of methods separating independent sources using fourth-order cumulants. Consider the case where the sources are independent, and the sum in the definition of J is taken only for $i = j$. Then we have a sum of the form $\sum_i \operatorname{cum}(\mathbf{w}_i^T \mathbf{z}(t), \mathbf{w}_i^T \mathbf{z}(t), \mathbf{w}_i^T \mathbf{z}(t-\Delta t), \mathbf{w}_i^T \mathbf{z}(t-\Delta t))$. (For temporally white data this cumulant is equal to the covariance $\operatorname{cov}([\mathbf{w}_j^T \mathbf{z}(t)]^2, [\mathbf{w}_i^T \mathbf{z}(t-\Delta t)]^2)$, see [10].) In the case where the lag Δt is taken zero, these cumulants are the kurtoses, and we see the connection to maximization of the squares (or, possibly, the absolute values) of kurtoses [4]. In the case of lagged cumulants (i.e. $\Delta t \neq 0$ as in our Theorem), it was proven in [10] that maximization of the square (or absolute value) of such

[1] Strictly speaking, the data should be temporally uncorrelated if it is generated according to the model defined here, so this temporal filtering should not be needed. However, since the model is necessarily only an approximation, it is useful to transform the data so that the approximation becomes better. Thus, temporal decorrelation is a useful practical procedure that is not necessary in the theoretical analysis, and is therefore not mentioned in the Theorem.

a cumulant leads to separation of one independent source, if it has a smoothly changing, nonstationary variance.

Likewise, the condition of full rank of $\mathbf{K}$ can be viewed as an extension of the classic condition of nonzero kurtosis [4,5], or, in general, the condition of nonzero fourth-order lagged cumulants [10]. In the case of independent sources, $\mathbf{K}$ is diagonal with the fourth-order cumulants $\mathrm{cum}(s_i(t), s_i(t), s_i(t-\Delta t), s_i(t-\Delta t))$ in its diagonal. Then, the assumption of full rank boils down to assuming that these lagged cumulants are non-zero, which reflects the assumption that the sources have time-dependencies due to smoothly changing variance variables $v_i(t)$. If the dependencies between the sources are weak enough that $\mathbf{K}$ is strictly diagonally dominant, $\mathbf{K}$ is of full rank as well [7]. However, the condition of full rank does not require, in general, that the dependencies are weak. They can be very strong, as strong off-diagonal elements have little to do with a matrix being singular.

In fact, the condition of full rank is violated, for example, if all the signals v_i, v_j are equal to each other for any i, j, and the y_i have identical distributions (all the entries in the matrix $\mathbf{K}$ are then equal). So, one intuitive consequence of the condition is that the variance signals v_i must have different time courses. They can have identical distributions over time, but they must not have identical time courses. For example, bursts of activity should be slightly delayed with respect to each other.

Strictly speaking, the variances $v_i(t)$ need not have autocorrelations, i.e. change smoothly in time. The matrix $\mathbf{K}$ can have full rank even if its diagonal is zero, which corresponds to the case where these autocorrelations are zero. However, the condition then requires that the signals $v_i(t)$ have nonzero crosscorrelations with time lag Δt, and it is difficult to imagine a real situation where the signals have nonzero crosscorrelations but zero autocorrelations.

Maximization of the contrast function can be performed by gradient ascent, where the new separating matrix is projected on the set of orthogonal matrices [14]; the gradient can also be first projected on the tangent space [6], which was done in the simulations below. Further improvements might be obtained by using a line search method as in [9], or conjugate gradients. A sketch of a Matlab implementation of an algorithm is given in Appendix B. In the typical case where the number of data points is much larger than the dimension of the data, the computational complexity of computing the gradient can be easily seen to be twice as large as in basic ICA. Yet, the actual computational complexity depends also on the number of iterations neeeded, and comparing them is difficult.

Generalization of the method to several time lags is straightforward. Denote by $J_{\Delta t}$ the objective function for the lag Δt. We can simply take the sum of the objective functions for several time lags $\Delta t(1), \ldots, \Delta t(k)$, and maximize the sum $\sum_{i=1}^{k} J_{\Delta t(i)}$.

4 Simulations

We performed simulations in an attempt to confirm the theoretical results above. The simulations consisted of 100 source separation trials with three different methods: 1) the double-blind method proposed in this paper, 2) FastICA using kurtosis [15], and 3) the method based on nonstationary variance proposed in [10] and closely related to those one in [19,17]. The methods in points 2) and 3) are the closest to the double-blind method in the set of blind source separation methods based on independence.

In each trial, we created five random signals of length 10,000 time points. First, we created the variance signals with the following method. Five time signals were created using a multivariate gaussian first-order autoregressive model. The matrix defining the AR(1) model was generated randomly in each trial, with gaussian coefficients. Outliers, defined as values larger than a threshold of 3 times the standard deviation, were eliminated from the resulting signals by reducing their values to the above-mentioned threshold (see below for a discussion of nonrobustness). The variance signals v_i were then defined as the absolute values of these signals. This gave variance signals that had strong correlations both over time and with each other, but no really large values that could lead to annoying outliers in the source signals. This latter point is important because cumulant-based methods are quite vulnerable to outliers.

Next, the source signals s_i were created by multiplying the variance signals by i.i.d. (white) zero-mean subgaussian random processes y_i, as in Eq. (2). The signals y_i had to be strongly subgaussian (here, signed fourth root of zero-mean uniform variables) because otherwise this construction does not create enough dependencies, and estimation is too easy for any method. The source signals were normalized to unit variance; they had zero mean by construction. Finally, a random mixing matrix $\mathbf{A}$ was created, and the signals were mixed to give the observed signals $x_i, i = 1, \ldots, 5$.

The three methods were then applied on the data. The performance of each method was assessed as follows. Denoting by $\mathbf{W}$ the obtained estimate of the inverse of the mixing matrix (with permutation and sign indeterminacies), we looked at the matrix $\mathbf{WA}$. We computed how many elements in this matrix had an absolute value that was larger than 0.99. This gave a measure of how many source signals had been separated. First of all, it must be noted that the matrix $\mathbf{WA}$ is rather exactly orthogonal (up to insignificant errors occurred in the estimation of the whitening matrix), so there can be no more than 5 such elements in the matrix, and no row or column can contain more than one such element. In the ideal case where $\mathbf{WA}$ is a signed permutation matrix, there would be exactly five such elements. Thus, this is a valid measure of the number of source signals separated.

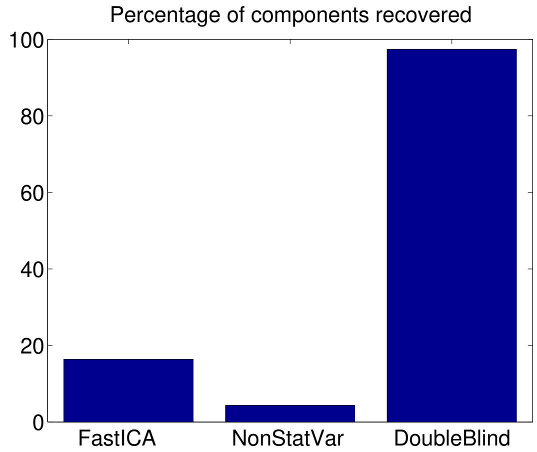

Fig. 2. Percentage of separated components by two conventional independence-based blind source separation methods (FastICA using kurtosis [15] and the nonstationary-variance-based algorithm in [10]), and our double-blind method.

The results are shown in Figure 2. Our method separated 97.4% of the components, whereas the other two methods separated less than 20% of the components. Thus, while not being perfect, our method was quite good, while the conventional independence-based methods performed miserably.

5 Discussion

Instead of using higher-order statistics, some methods separate signals by using the temporal second-order correlations [20,18,2]. It is sometimes claimed that these methods separate signals without assuming independence, only uncorrelatedness. It must be noted, however, that these methods also need to assume that the signals have different spectral characteristics, that is, different autocorrelation structures. Thus, the second-order methods have a considerably more limited domain of application, since in many practical cases, one wants to separate signals which have almost exactly the same characteristics.

Our double-blind method is clearly a very rudimentary one, and better methods should be developed. First, the method is very sensitive to outliers due to the use of fourth-order cumulants; more robust methods are needed in many applications. Second, it would be interesting to relax the assumption of temporal correlation, so that no temporal structure of signals is needed.

To conclude, we have proposed a framework for separating source signals that are dependent through their variances, corresponding to general activity levels. We assumed that the source signals have the same kind of temporal dependencies as well, that is, they have nonstationary smoothly changing variances. This made it possible to propose a cumulant-based contrast function that was shown to separate the signals without necessitating estimation of a model of the source dependencies.

A Proof of theorem

Now we prove the Theorem announced above. Denote $\mathbf{q}_i^T = \mathbf{w}_i^T \mathbf{A}$. Then $\mathbf{w}_i^T \mathbf{z}(t) = \mathbf{q}_i^T \mathbf{s}(t)$. Consider the cumulant

$$\tilde{k}_{ij} = \mathrm{cum}(\mathbf{q}_i^T \mathbf{s}(t), \mathbf{q}_i^T \mathbf{s}(t), \mathbf{q}_j^T \mathbf{s}(t-\Delta t), \mathbf{q}_j^T \mathbf{s}(t-\Delta t)) \tag{A.1}$$

Due to temporal uncorrelatedness, this equals [10] the squared covariance used in the Theorem:

$$\tilde{k}_{ij} = \mathrm{cov}([\mathbf{q}_i^T \mathbf{s}(t)]^2, [\mathbf{q}_j^T \mathbf{s}(t-\Delta t)]^2) \tag{A.2}$$

By the basic properties of cumulants [14] we have

$$\tilde{k}_{ij} = \sum_{klmn} q_{ik} q_{il} q_{jm} q_{jn} \mathrm{cum}(s_k(t), s_l(t), s_m(t-\Delta t), s_n(t-\Delta t)) \tag{A.3}$$

Now, the essential point is that all the cumulants of the form $\mathrm{cum}(s_k(t), s_l(t), s_m(t-\Delta t), s_n(t-\Delta t))$ are zero unless $k=l$ and $m=n$. This is because of the relation $s_i(t) = v_i(t) y_i(t)$, where $y_i(t)$ is independent from any $y_i(\tau), \tau \neq t$ and from any $y_j(t), j \neq i$, as well as from any $v_j(t)$. Consider, for example, the case where we have the constraints $k=m$ and $l=n$ instead. Then, by the well-known formula for the fourth-order cross-cumulant of zero-mean variables [14], we have

$$\begin{aligned}
&\mathrm{cum}(s_k(t), s_l(t), s_k(t-\Delta t), s_l(t-\Delta t)) \\
&\quad = E\{v_k(t)v_l(t)v_k(t-\Delta t)v_l(t-\Delta t)\}E\{y_k(t)y_l(t)y_k(t-\Delta t)y_l(t-\Delta t)\} \\
&\quad - E\{v_k(t)v_l(t)\}E\{v_k(t-\Delta t)v_l(t-\Delta t)\}E\{y_k(t)y_l(t)\}E\{y_k(t-\Delta t)y_l(t-\Delta t)\} \\
&\quad - E\{v_k(t)v_k(t-\Delta t)\}E\{v_l(t)v_l(t-\Delta t)\}E\{y_k(t)y_k(t-\Delta t)\}E\{y_l(t)y_l(t-\Delta t)\} \\
&\quad - E\{v_k(t)v_l(t-\Delta t)\}E\{v_k(t)v_l(t-\Delta t)\}E\{y_k(t)y_l(t-\Delta t)\}E\{y_k(t)y_l(t-\Delta t)\}
\end{aligned} \tag{A.4}$$

Now, the random variables $y_k(t), y_l(t), y_k(t-\Delta t), y_l(t-\Delta t)$ are mutually independent and zero-mean. Every term in the above cumulant has the expectation of a product which contains exactly one occurrence of either two or four of these random variables. The expectation of such a product is thus zero.

All other cases of the equalities between indices can be shown to give zero cumulants in the same way. The only exception is the case $k=l, m=n$, because then we have repetition of the two terms $y_k(t)$ and $y_m(t-\Delta t)$, and thus the expectation of a square, which is not zero. Note that this is true only for a non-zero lag Δt; this is why we must assume that the data has a temporal structure.

Thus, we have

$$\tilde{k}_{ij} = \sum_{kl} q_{ik}^2 q_{jl}^2 \mathrm{cum}(s_k(t), s_k(t), s_l(t-\Delta t), s_l(t-\Delta t))$$
$$= \sum_{kl} q_{ik}^2 q_{jl}^2 \mathrm{cov}(s_k^2(t), s_l^2(t-\Delta t)) \quad \text{(A.5)}$$

Denote by $\mathbf{Q} = \mathbf{WA}$ the matrix with $\mathbf{q}_i^T$ as rows, and by $\bar{\mathbf{Q}}$ the matrix obtained by raising each element of $\mathbf{Q}$ to the power of two. Then, the matrix $\tilde{\mathbf{K}}$ with elements $\tilde{k}_{ij}$ can be expressed as

$$\tilde{\mathbf{K}} = \bar{\mathbf{Q}}\mathbf{K}\bar{\mathbf{Q}}^T \quad \text{(A.6)}$$

which shows the remarkable phenomenon that the four-dimensional cumulant tensor is reduced to a simple two-dimensional matrix. This is partly due to our assumptions on the dependency structure, and partly due to the choice of the particular cumulants.

The objective function in the Theorem is the square of the Frobenius norm of $\tilde{\mathbf{K}}$, i.e. the sum of squares of the elements. Now, we need the following lemma, reminiscent of Lemma 15 in [4]:

Lemma 1 *Consider a matrix $\bar{\mathbf{Q}}$ that is doubly stochastic, i.e. the sums of rows and the sums of columns are all equal to one. Take any square matrix of the same dimensions $\mathbf{M}$ that has full rank. Then for the Frobenius norm it holds:*

$$||\bar{\mathbf{Q}}\mathbf{M}||^2 \leq ||\mathbf{M}||^2 \quad \text{(A.7)}$$

with equality if and only if $\bar{\mathbf{Q}}$ is a permutation matrix.

Proof of Lemma: According to a theorem by Birkhoff [7, p. 527], we can represent a doubly stochastic matrix as a finite convex sum of permutation matrices:

$$\bar{\mathbf{Q}} = \sum_s \alpha_s \mathbf{P}_s \quad \text{(A.8)}$$

with $\alpha_s > 0$ and $\sum_s \alpha_s = 1$. The converse also holds. The set of doubly stochastic matrices is thus a compact convex set with extreme points $\mathbf{P}_s$. On the other hand, the square of the Frobenius norm $||\bar{\mathbf{Q}}\mathbf{M}||^2$ is a strictly convex function of $\bar{\mathbf{Q}}$ (because $||\bar{\mathbf{Q}}||^2$ is trivially strictly convex, and a non-singular linear transformation does not change convexity). Thus, the maxima are obtained at the extreme points, i.e. when $\bar{\mathbf{Q}}$ is a permutation matrix, which proves the lemma.

Now, $\bar{\mathbf{Q}}$ in (A.6) is doubly stochastic since it consists of the squares of an orthogonal matrix, and $\mathbf{K}$ is assumed to have full rank. Applying the Lemma twice, we see that the (square of the) Frobenius norm of $\tilde{\mathbf{K}}$ is maximized exactly when $\bar{\mathbf{Q}}$ is a permutation matrix. This means that $\mathbf{Q} = \mathbf{WA}$ is a signed permutation matrix, and the sources have been separated. Thus, the theorem is proven.

B Algorithm for maximizing the objective function

Here we show how to code one step of an iterative algorithm for the maximixation of the objective function in the Theorem, in Matlab code.

Denote by T the number of time points. Denote by Z a $n \times (T-2)$ matrix that contains each $\mathbf{z}(t), t = 2 \ldots T-1$ as a column (the index begins with 2 to accomodate the lagged version below). Denote by Zplus a matrix that contains the lagged data, i.e. $\mathbf{z}(t), t = 1 \ldots T-2$, if the lag Δt is equal to 1, and likewise for Zminus that contains the "anti-lagged" data $\mathbf{z}(t), t = 3 \ldots T$. If the lag is different from 1, only the definitions of Zplus and Zminus need to be changed. The basic code is as follows:

```
%compute estimates of sources with lags
Y=W*Z;
Yplus=W*Zplus;
Yminus=W*Zminus;

%compute cumulant matrix
K=(Y.^2)*(Yminus'.^2)/T-mean(Y'.^2)'*mean(Yminus'.^2);

%compute gradient
KKtsum=diag(sum(K+K'));
grad=(Y.*(K*(Yminus.^2)+K'*(Yplus.^2)))*Z'/T-KKtsum*W;

%compute projection of gradient to the tangent plane
%of constraint surface (optional)
ortgrad=grad-W*grad'*W;

%do gradient step with some stepsize
%(could also use grad directly but this is better)
W=W+stepsize*ortgrad;

%project back to the constaint surface, i.e. orthogonalize
W=inv(sqrtm(W*W'))*W;
```

References

[1] F. R. Bach and M. I. Jordan. Tree-dependent component analysis. In *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI-2002)*, 2002.

[2] A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. on Signal Processing*, 45(2):434–444, 1997.

[3] J.-F. Cardoso. Multidimensional independent component analysis. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'98)*, Seattle, WA, 1998.

[4] P. Comon. Independent component analysis—a new concept? *Signal Processing*, 36:287–314, 1994.

[5] N. Delfosse and P. Loubaton. Adaptive blind separation of independent sources: a deflation approach. *Signal Processing*, 45:59–83, 1995.

[6] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[7] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[8] J. Hurri and A. Hyvärinen. Temporal and spatiotemporal coherence in simple-cell responses: A generative model of natural image sequences. Submitted manuscript.

[9] J. Hurri and A. Hyvärinen. Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Computation*, 15(3):663–691, 2003.

[10] A. Hyvärinen. Blind source separation by nonstationarity of variance: A cumulant-based approach. *IEEE Transactions on Neural Networks*, 12(6):1471–1474, 2001.

[11] A. Hyvärinen and P. O. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

[12] A. Hyvärinen, P. O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7):1527–1558, 2001.

[13] A. Hyvärinen, J. Hurri, and J. Väyrynen. Bubbles: A unifying framework for low-level statistical properties of natural image sequences. *J. of the Optical Society of America A*, 20(7), 2003. In press.

[14] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley Interscience, 2001.

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

[16] C. Jutten and J. Hérault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[17] K. Matsuoka, M. Ohya, and M. Kawamoto. A neural net for blind separation of nonstationary signals. *Neural Networks*, 8(3):411–419, 1995.

[18] L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72:3634–3636, 1994.

[19] D.-T. Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of non-stationary sources. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pages 187–193, Helsinki, Finland, 2000.

[20] L. Tong, R.-W. Liu, V.C. Soon, and Y.-F. Huang. Indeterminacy and identifiability of blind identification. *IEEE Trans. on Circuits and Systems*, 38:499–509, 1991.

[21] H. Valpola, M. Harva, and J. Karhunen. Hierarchical models of variance sources. In *Proc. Int. Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, 2003.

[22] R. Vigário, V. Jousmäki, M. Hämäläinen, R. Hari, and E. Oja. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. In *Advances in Neural Information Processing Systems*, volume 10, pages 229–235. MIT Press, 1998.

[23] M. J. Wainwright, E. Simoncelli, and A. S. Willsky. Random cascades on wavelet trees and their use in analyzing and modeling natural images. *Applied Computational and Harmonic Analysis*, 11:89–123, 2001.