BLIND SOURCE SEPARATION BASED ON SPACE-TIME-FREQUENCY DIVERSITY

Scott Rickard, Radu Balan, Justinian Rosca

Siemens Corporate Research, 755 College Road East, Princeton, NJ 08540
{scott.rickard,radu.balan,justinian.rosca}@scr.siemens.com

ABSTRACT

We investigate the assumption that sources have disjoint support in the time domain, time-frequency domain, or frequency domain. We call such signals disjoint orthogonal. The class of signals that approximately satisfies this assumption includes many synthetic signals, music and speech, as well as some biological signals. We measure the disjoint orthogonality of the benchmark signals in the ICALAB Toolbox in the time, time-frequency, and frequency domains and show that most satisfy the assumption in at least one representation. In order to compare this assumption with other common source assumptions, we derive a demixing algorithm for noisy instantaneous mixtures based on disjoint orthogonality and compare its performance to the algorithms in the ICALAB Toolbox, all of which rely on the second-order statistics, non-stationarity, or higher-order statistics of the sources. The results indicate that space-time-frequency diversity is a useful assumption for the design of BSS/ICA algorithms. [1]

1. INTRODUCTION

Blind source separation and independent component analysis algorithms leverage the knowledge that the sources satisfy certain statistical or deterministic conditions in order to perform the separation. Amari and Chiocki [1] list four common source property assumptions that form the basis for most BSS/ICA algorithms:

1. Higher-order statistics (HOS). Sources are statistically independent. This is usually practically enforced by looking to the 4th order moments or cumulants of the mixtures.
2. Second-order statistics (SOS). Sources are decorrelated.
3. Non-stationarity and SOS (NS). Sources are decorrelated and have time-varying variances.
4. Space-time-frequency diversity (STF). Sources are disjoint in the time domain, time-frequency domain, or frequency domain. This is the assumption we analyze in this paper. An alternative STF assumption is presented in [2].

Most methods fall into one of the first three categories, and few techniques make use of space-time-frequency diversity. For example, the ICALAB Toolbox [3], a software program that allows one to compare the performance of BSS/ICA algorithms, contains 19 BSS/ICA methods, none of which can be classified as a STF method.

Surprisingly, however, of the 17 benchmark signal families contained in the ICALAB Toolbox, we will show that 15 of them possess a large degree of time-frequency diversity. Moreover, none of the techniques (including the one we present here) are able to demix the two benchmark families which do not possess a high level of time-frequency diversity. So, all the practically "demixable" fifteen benchmarks are time-frequency diverse.

Specifically, in this paper we analyze the assumption that the sources have disjoint support in either the time domain, frequency domain, or the time-frequency domain. We call signals for which there exists an invertible linear transform such that in the transform domain the signals have disjoint support *disjoint orthogonal*. When the transform is the windowed Fourier transform, we call the signals *W-disjoint orthogonal*. For such disjoint orthogonal sources, we derive a separation algorithm.

While the disjoint orthogonal source assumption may seem too restricting, we argue that it is in practice approximately satisfied by many signals of interest. Specifically, time-division multiplexed communication signals are by design time domain disjoint, frequency-division multiplexed communication signals are by design frequency domain disjoint, and the goal of frequency hopped CDMA signals is that the signals are disjoint in the time-frequency domain. Additionally, perhaps surprisingly, speech signals are W-disjoint orthogonal enough to allow for accurate mixing parameter estimation and blind separation [4]. Indeed, as we will show, music and speech, as well as some biological signals, are approximately W-disjoint orthogonal.

For the separation algorithm, we consider an additive noise mixing model with an arbitrary number of sensors and possibly more sources than sensors (the "degenerate separation problem"). The basis for our approach to noisy model estimation by maximum likelihood, under the instantaneous mixing assumptions, is that the sources are disjoint orthogonal. The implementation of the derived criterion involves iterating two steps: a partitioning of the time-frequency plane for separation followed by an optimization of the mixing parameter estimates. The solution is applicable to an arbitrary number of sensors and sources. That is, one can demix by converting the partitioned time-frequency representations back into the time domain. However, in order to compare with the other methods in the ICALAB Toolbox, we will use the estimate of the mixing matrix and perform standard inverse mixing matrix demixing. Experimentally, we show the capability of the technique on the ICALAB data.

The organization of the paper is as follows. Section 2 presents the signal mixing model and Section 3 provides experimental motivation of the W-disjoint orthogonality signal model. Section 4 shows the derivation of the ML estimator of mixing parameters and source signals, and its implementation by an iterative procedure. The algorithm performance on the ICALAB benchmarks is compared to the performance of the other ICALAB techniques in Section 5.

Scott Rickard is also with the Program in Applied and Computational Mathematics, Princeton University.

[1] In Proceedings of the 4th ICA-BSS Conference, April 2003, Japan

2. MIXING MODEL AND SIGNAL ASSUMPTION

2.1. The Mixing Model

Consider the measurements of L source signals by D sensors in an instantaneous mixing model:

$$\begin{bmatrix} x_1(t) \\ \vdots \\ x_D(t) \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1L} \\ \vdots & \ddots & \vdots \\ a_{D1} & \cdots & a_{DL} \end{bmatrix} \begin{bmatrix} s_1(t) \\ \vdots \\ s_L(t) \end{bmatrix} + \begin{bmatrix} n_1(t) \\ \vdots \\ n_D(t) \end{bmatrix} \quad (1)$$

where $A = (a_{dl})_{1\le d\le D, 1\le l\le L}$ is the instantaneous mixing matrix. We assume A has full rank, which ensures "space" diversity.

We denote by $X_d(k,\omega)$, $S_l(k,\omega)$, $N_d(k,\omega)$ the windowed Fourier transform of signals $x_d(t), s_l(t)$, and $n_d(t)$, respectively, with respect to a window $W(t)$, where k is the frame index, and ω the frequency index. When no danger of confusion, we shall drop the arguments k,ω in X_d, S_l and N_d. The mixing model (1) is thus

$$X_d(k,\omega) = \sum_{l=1}^{L} a_{dl} S_l(k,\omega) + N_d(k,\omega)\ ,\ 1 \le d \le D \quad (2)$$

or, more compactly,

$$X(k,\omega) = AS_{tot}(k,\omega) + N(k,\omega) \quad (3)$$

where X, N are the D-vectors of components X_d, N_d, and S_{tot} is the L-vector of components S_l. We shall denote by A_l the lth column of A, $A = [A_1|\cdots|A_L]$.

Our problem is: given measurements $(x_1(t), \ldots, x_D(t))_{1\le t\le T}$ we want to determine the ML estimates of the mixing parameters $A = (a_{dl})$ and the source signals $(s_1(t), \ldots, s_L(t))_{1\le t\le T}$. In order to solve this we rely on the W-disjoint orthogonality of the sources and the assumption that the sensor noises are independently distributed and have Gaussian distributions with zero mean and σ^2 variance.

2.2. The W-Disjoint Orthogonal Signal Model

In [5] we called two signals s_1 and s_2 *W-disjoint orthogonal* (W-DO), for a given windowing function $W(t)$, if the supports of the windowed Fourier transforms of s_1 and s_2 are disjoint, that is:

$$S_1(k,\omega)S_2(k,\omega) = 0\ ,\ \forall k,\omega \quad (4)$$

For L sources $S_1,\ldots,S_L$ the assumption generalizes to:

$$S_i(k,\omega)S_j(k,\omega) = 0\ ,\ \forall\, 1 \le i \ne j \le L,\ \forall k,\omega \quad (5)$$

Such a deterministic constraint is not only rarely satisfied, but it also implies that the signals are, in general, statistically dependent, which is easily proved by the fact that the conditional distribution $p(S_1 = s_1|S_2 \ne 0) = \delta(s_1)$ is different from the conditional $p(S_1 = s_1|S_2 = 0)$. In [6] it has been noticed, however, that relation (4) is satisfied in an approximate sense by real speech signals. Thus, (4) can be seen as the mathematical idealization of the condition that each mixture time-frequency point with significant power is most often dominated by a single source. The case of single source dominance of speech mixtures in the time-frequency domain has been noticed and utilized several times [7, 8, 9, 10, 11, 12]. It was also shown in [13] that (4) is the limit of a stochastic source model.

3. W-DISJOINT ORTHOGONALITY OF THE BENCHMARKS

We proposed in [6] the normalized difference between the signal energy contained in the dominant time-frequency points of a signal in a mixture and the interference energy in those points as a measure of W-disjoint orthogonality. In order to measure W-disjoint orthogonality for a signal of interest in a mixture for a given representation, we partition the points (in the TD, TF, or FD representation) of the mixture into those dominated by the source of interest and those dominated by the interference. We define the mask which is the indicator function of the *dominant* time-frequency points for source j

$$\Phi_j(k,\omega) = \begin{cases} 1 & |S_j(k,\omega)| > |Y_j(k,\omega)| \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where Y_j is the interference,

$$Y_j(k,\omega) = \sum_{\substack{i=1 \\ i\ne j}}^{N} S_i(k,\omega). \quad (7)$$

Now we define two important performance criteria: (1) how well the dominant time-frequency points preserve the source of interest, and (2) how well considering only the dominant time-frequency points suppresses the interfering sources. We define the preserved-signal-ratio (PSR) of a source in a mixture as

$$\text{PSR}(j) = \frac{\|\Phi_j(k,\omega)S_j(k,\omega)\|^2}{\|S_j(k,\omega)\|^2} \quad (8)$$

which measures the percentage of energy of source j contained in its dominant time-frequency points. We define the signal-to-interference ratio of the dominant time-frequency points of a source in a mixture,

$$\text{SIR}(j) = \frac{\|\Phi_j(k,\omega)S_j(k,\omega)\|^2}{\|\Phi_j(k,\omega)Y_j(k,\omega)\|^2}. \quad (9)$$

These two criteria, the PSR and SIR, are combined to form the measure of W-DO.

$$\begin{aligned} \text{WDO}(j) &= \frac{\|\Phi_j(k,\omega)S_j(k,\omega)\|^2 - \|\Phi_j(k,\omega)Y_j(k,\omega)\|^2}{\|S_j(k,\omega)\|^2} && (10) \\ &= \text{PSR}(j) - \text{PSR}(j)/\text{SIR}(j). && (11) \end{aligned}$$

For signals which have disjoint support, we note that PSR $= 1$, SIR $= \infty$, and thus WDO $= 1$. Moreover, WDO $= 1$ implies that PSR $= 1$, SIR $= \infty$, and that the signal has disjoint support compared to the interfering sources. In general, $0 \le$ PSR ≤ 1, SIR ≥ 0, and WDO ≤ 1 (and can be negative).

In order to summarize the W-disjoint orthogonality of a family of signals, we look to the average WDO and the minimum WDO, defined as follows,

$$\text{aWDO} = \frac{1}{L}\sum_{j=1}^{L} \text{WDO}(j) \quad (12)$$

$$\text{mWDO} = \min_j \text{WDO}(j) \quad (13)$$

Figure 1 lists the 17 benchmark signal families from the ICALAB Signal Processing Toolbox (Version 1.1) [3]. Benchmarks ACsin10d, ACvsparse10, and Speech10 are displayed in Figure 2. We measure the W-disjoint orthogonality of the benchmarks for three window sizes: one sample, 512 samples, and the signal length. These three sizes correspond to the time domain (TD), time-frequency domain (TF), and frequency domain (FD) representations of the signal. In the TD case, the ω frequency index is meaningless and in the FD representation the k time index is meaningless. However, for ease of notation and reference, we will use (k, ω) and refer to "time-frequency points" for all three representations. The average and minimum WDO for the ICALAB benchmarks are listed in Figure 3 for the three representations of the signals. Note that all but 2 (AC10-7sparse and EEG19) have equal to or greater than 40% average WDO. For each signal, the largest aWDO and mWDO is highlighted in bold.

ACsin10d	10 sine waves
ACsin4d	4 sine waves
ACsparse10	10 sparse bell-shaped sources
ACvsparse10	10 very sparse spiking signals
ABio7	7 typical biological signals
Sergio7	7 random sources, some asymmetrically distributed
AC10-7sparse	10 sources mixtures of 7 from ACsparse10
acspeech16	16 typical speech signals
Speech4	4 speech and music sources
Speech8	8 speech and music sources
Speech10	10 speech and music sources
Speech20	20 speech and music sources
10halo	10 speakers saying the same thing
20depspeech	20 speakers saying the same thing
nband5	5 narrow band sources
Gnband	5 forth order colored sources
EEG19	19 EEG signals

Fig. 1. ICALAB benchmarks.

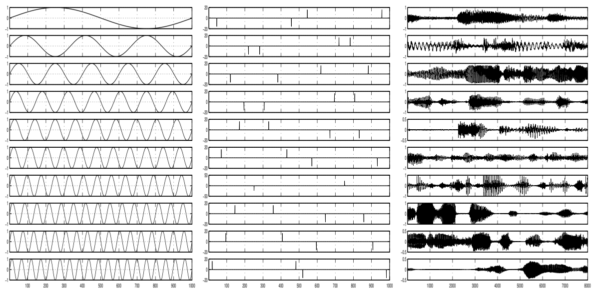

Fig. 2. Benchmarks ACsin10d, ACvsparse10, and Speech10.

	TD		TF		FD	
benchmark	aWDO	mWDO	aWDO	mWDO	aWDO	mWDO
ACsin10d	0.26	0.16	0.46	0.38	**0.82**	**0.73**
ACsin4d	0.29	0.06	**1.00**	**1.00**	0.99	0.99
ACsparse10	**0.69**	**0.45**	0.24	0.13	0.22	0.09
ACvsparse10	**1.00**	**1.00**	0.25	0.12	0.21	0.12
ABio7	0.41	0.27	0.52	0.39	**0.64**	**0.50**
Sergio7	0.32	0.09	**0.45**	0.11	0.35	**0.11**
AC10-7sparse	**0.03**	**0.00**	0.02	0.00	0.00	0.00
acspeech16	0.26	0.22	**0.59**	**0.38**	0.48	0.32
Speech4	0.59	0.42	**0.87**	**0.81**	0.79	0.73
Speech8	0.35	0.18	**0.68**	**0.49**	0.45	0.21
Speech10	0.31	0.15	**0.64**	**0.43**	0.41	0.19
Speech20	0.24	0.19	**0.60**	**0.42**	0.49	0.32
10halo	0.34	0.28	**0.52**	**0.42**	0.39	0.29
20depspeech	0.24	0.14	**0.37**	**0.22**	0.27	0.17
nband5	0.37	0.35	**0.99**	**0.98**	0.99	0.98
Gnband	0.37	0.34	**0.40**	0.36	0.39	**0.38**
EEG19	**0.06**	0.00	0.03	**0.01**	0.03	0.00

Fig. 3. Average and minimum WDO for the ICALAB benchmarks.

Most of the sources exhibit a high level of disjoint orthogonality, but we need to, given only the mixtures, determine in which representation (TD, TF, or FD) the sources are most W-DO. One approach would be to run the algorithm described in the next section three times, once in each domain and then choose the solution with the most W-DO outputs. Alternatively, we can measure how sparse the mixtures are in each representation and then run the algorithm on the most sparse representation. The logic in this is that the disjoint orthogonality comes from, in general, each signal having a sparse representation and the few large coefficients of each source not overlapping with one another. Because of linearity, sparse signal representations should lead to sparse mixture representations and we hope that the most sparse mixture representation corresponds to the representation with the largest W-DO. In Appendix A, we show that this logic holds true for the speech and music benchmarks and some of the synthetic benchmarks, but fails for some of the synthetic benchmarks.

4. THE MAXIMUM LIKELIHOOD ESTIMATOR OF SIGNAL AND MIXING PARAMETERS

In this section we derive the joint maximum likelihood estimator of parameters and source signals under assumption (5). The source signals naturally partition the time-frequency plane into L disjoint subsets $\Omega_1, \ldots, \Omega_L$, where each source signal is non-zero (i.e. active). Thus the signals are given by the collection $\Omega_1,\ldots,\Omega_L$ and one complex variable S that defines the active signal:

$$S_l(k,\omega) = S(k,\omega) 1_{\Omega_l}(k,\omega) \tag{14}$$

Let the model parameters θ consist of the mixing parameters $A = (a_{dl})_{1\le d\le D, 1\le l\le L}$, the partition $(\Omega_l)_{1\le l\le L}$ and S. Its likelihood and maximum log-likelihood estimator are given by:

$$\mathcal{L}(\theta) = \prod_{l=1}^{L} \prod_{(k,\omega)\in\Omega_l} \frac{1}{(\pi\sigma^2)^D} \exp\{-\frac{1}{\sigma^2}\|X(k,\omega) - A_l S(k,\omega)\|^2\}$$
$$\hat{\theta}_{ML} = \mathrm{argmin}_\theta \sum_{l=1}^{L} \sum_{(k,\omega)\in\Omega_l} \|X - A_l S\|^2 \tag{15}$$

For any partition $(\Omega_1, \ldots, \Omega_L)$ we define the selection map Σ : TF-plane $\to \{1, \ldots, L\}$, $\Sigma(k,\omega) = l$ iff $(k,\omega) \in \Omega_l$. Clearly Σ defines a unique partition. Optimizing over S in (15) we obtain

$$\hat{S} = \frac{A_l^* X}{\|A_l\|^2} \tag{16}$$

where $l = \Sigma(k,\omega)$. Inserting (16) into (15), the optimization problem reduces to:

$$(\hat{A}, \hat{\Sigma}) = \mathrm{argmax}_{A,\Sigma} J(A,\Sigma) \tag{17}$$

where:

$$J(A,\Sigma) = \sum_{(k,\omega)} \frac{|A^*_{\Sigma(k,\omega)} X(k,\omega)|^2}{\|A_{\Sigma(k,\omega)}\|^2} \tag{18}$$

Note the criterion to maximize depends on a set of continuous parameters A, and a selection map Σ. A typical optimization algorithm for such a criterion works as follows. The optimization is done in two steps: first the optimization over the continuous parameters, and then the optimization over the selection map (or, equivalently, the partition). Such a procedure is iterated until the criterion reaches a saturation floor. Because the criterion is bounded above, we are guaranteed it will converge. Next we describe solutions for the two optimization problems.

4.1 Optimal Partition

Given a set of mixing parameters, $A = (a_{dl})_{1\le d\le D, 1\le l\le L}$, the optimal selection map is simply given by

$$\hat{\Sigma}(k,\omega) = \mathrm{argmax}_l \frac{|A_l^* X(k,\omega)|^2}{\|A_l\|^2} \tag{19}$$

The partition is then immediate: $\Omega_l = \{(k,\omega) | \Sigma(k,\omega) = l\}$.

4.2 Optimal Mixing Parameters

Now given a partition $(\Omega_l)_{1\le l\le L}$, the optimal mixing parameters are obtained independently for each l by:

$$\hat{A}_l = \mathrm{argmax}_{A_l} \sum_{(k,\omega)\in\Omega_l} \frac{|A_l^* X(k,\omega)|^2}{\|A_l\|^2} \tag{20}$$

Denote

$$R_l = \sum_{(k,\omega)\in\Omega_l} X(k,\omega)X(k,\omega)^* \tag{21}$$

Then the optimization problem turns into:

$$\hat{A}_l = \text{argmax}_{A_l} \frac{A_l^* R_l A_l}{\|A_l\|^2} \tag{22}$$

whose solution is the main eigenvector of the symmetric and non-negative matrix R_l,

$$R_l \hat{A}_l = \lambda \hat{A}_l \ , \ \ \lambda \geq \mu \text{ s.t. } R_l Z = \mu Z \ , \text{ for some } Z \neq 0 \tag{23}$$

4.3 ML Algorithm

Summing these findings, the optimization algorithm becomes:

- Step 0. Initialize $A^0 = (a_{dl})_{1\leq d\leq D, 1\leq l\leq L}$ with, for instance random values (or with the method presented in Appendix B); Set $s = 0$, $J^s = 0$, and choose a stopping threshold ϵ;
- Step 1. Find the optimal partition $(\Omega_l^{s+1})_{1\leq l\leq L}$, and selection map, Σ^{s+1} by solving (19) with $A = A^s$;
- Step 2. Find the optimal parameters A_l^{s+1} from (23) for the partition $\Omega_l = \Omega_l^{s+1}$, and all $1 \leq l \leq L$;
- Step 3. Set $s = s + 1$, and compute $J^s = J(A^s, \Sigma^s)$. If $(J^s - J^{s-1})/J^s > \epsilon$ then go to Step 1; otherwise:
- Step 4. The exit values are $A = A^s$, and $\Omega_l = \Omega_l^s$, obtained after s iterations. The source signal are then computed by converting the estimated time-frequency representations back into the time domain.

The core of this algorithm is essentially the same as that presented in [9]. It can also be seen as a specification to instantaneous mixtures of the anechoic mixing method presented in [13].

4.4 BSS Algorithm (STF-ER)

In order to compare with the techniques in the ICALAB Toolbox, we do not demix via partitioning, but rather use the mixing matrix estimate and standard mixing matrix inversion demixing. The overall BSS algorithm based on space-time-frequency diversity which operates in the most efficient representation (STF-ER) is as follows:

- Measure the 95% efficiency as described in Appendix A and select the most efficient representation.
- For the mixtures in the most efficient representation, initialize the mixing matrix estimate by clustering a random selection of the time-frequency points which make up the 95% efficiency as described in Appendix B.
- Loop through Steps 1–3, measure the aWDO and mWDO of the estimated outputs after each mixing matrix reestimation.
- After convergence of the criterion (or a fixed number of loops), invert the mixing matrix estimate corresponding to the largest sum of aWDO and mWDO, and apply it to the mixtures to produce the original source estimates.

5. EXPERIMENTAL RESULTS

We first tested the STF-ER algorithm on square mixtures of the signals in 10 of the 17 ICALAB benchmarks. No algorithm was able to demix either AC10-7sparse or EEG19 because they both contain nearly identical source signals, so these benchmarks were eliminated from the test set. ACsin4d, Speech{4,8,20}, and 20depspeech were also not considered because other signals contained in the test set were very similar. For the remaining 10 signals, in order to show that the method presented here has the possibility of working, we initialized the method with the partition assigning each time-frequency point to the corresponding largest magnitude original source at that time-frequency point. For W-DO sources, this is the optimal partition. We then performed one Step 2 mixing matrix estimation and stopped. This non-blind algorithm was run in the time domain (STF-TD-OP), time-frequency domain (STF-TF-OP), and frequency domain (STF-FD-OP). STF-ER-OP selects the most efficient representation and then runs the appropriate method. The results, shown in Figure 5 demonstrate that partitions do exist which allow for demixing. More details concerning the Performance Index (PI) measure of demixing performance can be found in [1] and [3]. Lower PI scores are better, zero implies perfect demixing, and, for our purposes, a PI score less than 0.1 indicates good demixing performance.

For the comparison experiments we tested STF-ER against the 19 algorithms in ICALAB, which are listed in Figure 4. For the tests, all algorithms were run with the default parameter settings. For a detailed description and discussion of the algorithms consult [1] and [3]. For STF-ER, Steps 1–3 were looped 5 times and the entire algorithm was run 9 times with the demixing matrix producing the outputs with the largest sum of aWDO and mWDO being the one selected. We first tested the algorithms using the identity matrix as the mixing matrix. Results are presented in Figure 6. The purpose of these tests was to expose the algorithms with source assumptions inconsistent with the properties of the benchmarks. As the mixtures are already demixed, the correct algorithm behavior would be to leave them unaltered, but as the results indicate, this rarely happens. In fact, EVD24 is the only algorithm that has a PI of less than 0.1 (good demixing performance) for all benchmark files. SOBI, SOBI-RO, JADETD, SANG, and STF-ER all demix seven of the ten benchmarks. NG-FICA failed to demix (PI > 0.1) any of the benchmark files.

We also tested the algorithms on random square mixtures; The results are presented in Figure 7. STF-ER-OP demixed 9 of the 10 benchmarks consistently with ACsparse10 being the one benchmark it "failed" to demix. STF-ER-OP demixed ACsparse10 into a series of sinsoids instead of a number of time disjoint bell-shaped bumps. In fact, several of the techniques proposed this alternative demixing. On the other hand, several of the techniques which failed on ACsin10d did so because they proposed demixtures that looked like ACsparse10. JADETD and SANG both consistently demix 7 of the 10 benchmarks. STF-ER consistently demixes 6 of the 10 benchmarks, and is the only method to demix the time disjoint ACvsparse10 and the frequency disjoint ACsin10d. EVD24 (which demixed all benchmarks in the identity case) and NG-FICA both failed in this demixing test on all of the benchmark files.

AMUSE	Algorithm for Multiple Unknown Source Extraction based on EVD
BSS SVD	BSS SOS algorithm based on SVD
EVD2	BSS SOS algorithm based on symmetric EVD
SOBI	Second Order Blind Identification
SOBI-RO	Robust SOBI with Robust Orthogonalization
SOBI-BPF	Robust SOBI with bank of Band-Pass Filters
SONS	Second Order Nonstationary Source Separation
EVD24	BSS SOS+FOS algorithm based on symmetric EVD
JADEop	Robust Joint Approx. Diagonalization of Eigenmatrices with optimized numerical procedures
JADETD	HOS Joint Approximate Diagonalization of Eigen matrices with Time Delays
FPICA	Fixed-Point ICA
Pearson opt.	Pearson system optimized
SANG	Self Adaptive Natural Gradient algorithm with nonholonomic constraints
NG-FICA	Natural Gradient - Flexible ICA
NG-OL	On-line adaptive Natural Gradient
ERICA	Equivariant Robust ICA - based on Cumulants
SIMBEC	SIMultaneous Blind Extraction using Cumulants
UNICA	Unbiased quasi Newton algorithm for ICA
FOBI-E	Fourth Order Blind Identification with Transformation matrix E

Fig. 4. BSS/ICA Algorithms in ICALAB [3]. EVD = eigenvector decomposition. FOS = forth order statistics.

6. SUMMARY

We have investigated the assumption that sources have disjoint support in the time domain, time-frequency domain, or frequency domain as a basis for blind source separation. Tests on the benchmarks signals in the ICALAB Toolbox reveal that, perhaps surprisingly, most of them exhibit a large degree of W-disjoint orthogonality in at least one domain. Based on this assumption, we derived a blind separation algorithm and tested it using the ICALAB benchmarks. The results show that there exist partitions of the domain which result in near perfect mixing matrix estimation. Iterative blind estimation of the partition results in performance comparable to other established BSS/ICA methods. This disparity in potential performance and actual performance suggests that future work in space-time-frequency methods may produce extremely powerful blind source separation methods.

alg	ACsin10d	ACsparse10	ACvsparse	ABio7	Sergio	acspeech16	ACSpeech10	10halo	nband5	Gnband
STF-TD-OP	0.3758	**0.0207**	**0.0000**	0.0435	0.2417	0.0228	0.0264	0.0337	0.0196	0.0248
STF-TF-OP	0.0282	0.2145	0.1161	0.0344	**0.0104**	**0.0110**	**0.0112**	**0.0250**	**0.0005**	**0.0155**
STF-FD-OP	**0.0095**	0.2847	0.1304	**0.0286**	0.0270	0.0130	0.0195	**0.0250**	**0.0005**	**0.0155**
STF-ER-OP	0.0282	0.2847	**0.0000**	**0.0286**	0.0270	**0.0110**	**0.0112**	**0.0250**	**0.0005**	**0.0155**

Fig. 5. Demixing Performance Index (PI) for ICALAB benchmarks for identity matrix mixing given W-DO optimal partition.

alg	ACsin10d	ACsparse10	ACvsparse	ABio7	Sergio	acspeech16	ACSpeech10	10halo	nband5	Gnband
AMUSE	**0.0000**	0.6310	0.2839	**0.0395**	0.2691	**0.0283**	**0.0464**	**0.0777**	**0.0006**	0.5097
BSS SVD	**0.0000**	0.6314	0.2452	**0.0555**	0.1304	**0.0544**	**0.0478**	**0.0815**	**0.0027**	0.3493
EVD2	**0.0020**	0.6317	**0.0000**	**0.0605**	0.2091	**0.0301**	**0.0351**	0.1260	**0.0009**	0.3690
EVD24	**0.0000**	**0.0000**	**0.0000**	**0.0183**	**0.0030**	**0.0069**	**0.0074**	**0.0163**	**0.0005**	**0.0066**
SOBI	**0.0022**	0.6318	**0.0000**	**0.0235**	0.1412	**0.0142**	**0.0142**	**0.0477**	**0.0005**	0.3612
SOBI-RO	**0.0134**	0.6397	0.2146	**0.0413**	**0.0981**	**0.0244**	**0.0139**	**0.0310**	**0.0292**	0.2219
SOBI-BPF	**0.0081**	0.6337	0.1302	**0.0548**	0.1073	**0.0347**	**0.0213**	0.1097	**0.0293**	0.2561
SONS	**0.0085**	0.6398	0.3686*	0.1185	0.1598	**0.0299**	**0.0197**	**0.0355**	**0.0011**	0.3952
JADEop	0.6206	**0.0273**	**0.0458**	0.1312	**0.0108**	**0.0840**	**0.0237**	**0.0992**	0.3421	0.3332
JADETD	0.6110	**0.0551**	**0.0224**	**0.0836**	0.1646	**0.0758**	**0.0586**	0.1239	**0.0053**	**0.0292**
FPICA	0.3176	0.2547	**0.0000**	**0.0538**	**0.0051**	**0.0322**	**0.0395**	**0.0552**	0.1731	0.1635
Pearson opt.	0.6357	**0.0057**	**0.0000**	0.1434	0.1198	**0.0348**	**0.0338**	**0.0455**	0.3552	0.3253
SANG	0.6336	**0.0047**	**0.0000**	**0.0420**	**0.0039**	**0.0560**	**0.0357**	**0.0564**	0.1483	0.2194
NG-FICA	0.5556	0.5555	0.3333*	0.6667	0.5000	0.3333	0.7778	0.2222	0.2500	0.2500
NG-OL	0.3406	0.2260	0.1654	0.2272	**0.0057**	**0.0589**	**0.0310**	**0.0345**	0.3082	0.3053
ERICA	0.4062	0.2245	**0.0000**	0.2428	**0.0109**	0.1062	**0.0729**	0.1239	0.3930	0.3610
SIMBEC	0.2630	0.3154	**0.0000**	0.2346	**0.0058**	**0.0962**	**0.0187**	**0.0705**	0.2617	0.2891
UNICA	0.4189	0.2100	**0.0000**	0.2976	**0.0108**	0.1057	**0.0730**	0.1232	0.3619	0.4480
FOBI-E	0.4074	0.3187	**0.0320**	0.1625	**0.0726**	0.2103	0.2157	0.2672	0.4090	0.3428
STF-ER	**0.0086**	0.6253	**0.0000**	**0.0686**	0.3276	**0.0172**	**0.0194**	**0.0372**	**0.0006**	0.4438

Fig. 6. Demixing Performance Index (PI) for ICALAB benchmarks for identity matrix mixing. A '*' indicates that noise was added to avoid program execution error (Gaussian noise 20 dB SNR). Those with PI less than 0.1 are in bold to signify good demixing performance.

alg	ACsin10d	ACsparse10	ACvsparse	ABio7	Sergio	acspeech16	ACSpeech10	10halo	nband5	Gnband
STF-TD-OP	·	x	x	x	·	x	x	x	·	x
STF-TF-OP	x	·	·	x	x	x	x	x	x	x
STF-FD-OP	x	·	·	x	x	x	x	x	x	x
STF-ER-OP	x	·	x	x	x	x	x	x	x	x
AMUSE	x	·	·	x	·	x	x	x	x	·
BSS SVD	x	·	·	x	·	x	x	x	x	·
EVD2	x	·	·	x	·	x	x	·	x	·
EVD24	·	·	·	·	·	·	·	·	·	·
SOBI	x	·	·	x	·	x	x	·	x	·
SOBI-RO	x	·	·	x	·	x	x	x	·	·
SOBI-BPF	x	·	·	·	·	x	x	·	·	·
SONS	x	·	·	·	·	x	x	x	x	·
JADEop	·	x	x	·	·	·	x	x	·	·
JADETD	·	x	x	x	·	x	x	·	x	x
FPICA	·	·	x	x	x	x	x	x	·	·
Pearson opt.	·	x	x	·	x	x	x	x	·	·
SANG	·	x	x	x	x	x	x	x	·	·
NG-FICA	·	·	·	·	·	·	·	·	·	·
NG-OL	·	·	·	·	x	·	x	x	·	·
ERICA	·	·	x	·	x	·	x	·	·	·
SIMBEC	·	·	x	·	x	·	x	x	·	·
UNICA	·	·	x	·	x	·	x	·	·	·
FOBI-E	·	·	x	·	·	·	·	·	·	·
STF-ER	x	·	x	·	·	x	x	x	x	·

Fig. 7. Algorithm demixing performance for random mixtures. Each method was tested on five mixtures mixed with a randomly generated non-singular mixing matrix. A 'x' indicates that the PI score was less than 0.1 for all five tests. A · indicates the method failed to demix with a PI score less than 0.1 at least once in the five tests.

Appendix A - Sparseness Measure

We measure sparseness, the property that a small percentage of the signal coefficients (in either TD, TF, or FD) captures a large percentage of the signal energy, as follows. For a fixed β,

$$\alpha^*(\beta) = \text{argmax}_\alpha \sum_{(k,\omega):|X(k,\omega)|\geq\alpha} |X(k,\omega)|^2 \geq \beta \sum_{(k,\omega)} |X(k,\omega)|^2 \tag{24}$$

Then the β efficiency level is

$$\text{eff}(\beta) = \frac{|\{(k,\omega) : |X(k,\omega)| < \alpha^*(\beta)\}|}{|\{(k,\omega)\}|} \tag{25}$$

For example, the 95% efficiency would be the maximum percentage of components (clearly the smallest magnitude ones) that we can throw away while still maintaining at least 95% of the signal energy. A threshold independent measure of efficient is,

$$\text{sumeff} = \int_0^1 \text{eff}(\beta) d\beta \tag{26}$$

Figure 8 shows the 95% efficiency level and sumeff for the ICALAB benchmarks. Each benchmark was mixed using the identity matrix before the efficiencies were calculated. Comparing Figure 8 to Figure 3, we note that the most efficient representation corresponds to the maximum average WDO representation in 11 out of the 15 "demixable" benchmarks. The remaining four; ACsin10d, ACsin4d, Sergio7, and ACsparse10 are more efficiently represented in one domain but more W-DO in another. This difference is most pronounced in the case of ACsparse10, which has signals consisting of time disjoint bell-shaped bumps. Mixtures of ACsparse10 appear sinusoidal. Thus, the signals of ACsparse10 are significantly more W-DO in the time domain but the mixtures of ACsparse10 are more efficiently represented in the time-frequency domain.

	TD		TF		FD	
benchmark	95%	sumeff	95%	sumeff	95%	sumeff
ACsin10d	0.69	0.91	**0.98**	**0.99**	0.97	0.99
ACsin4d	0.50	0.83	0.97	0.99	**0.98**	**0.99**
ACsparse10	0.29	0.71	0.98	0.99	**0.98**	**1.00**
ACvsparse10	**0.96**	**0.98**	0.52	0.85	0.63	0.89
ABio7	0.44	0.82	0.66	0.94	**0.71**	**0.95**
Sergio7	0.43	0.82	0.52	0.88	**0.57**	**0.90**
AC10-7sparse	0.66	0.91	**0.98**	**0.99**	0.97	0.99
acspeech16	0.50	0.84	**0.59**	**0.90**	0.52	0.87
Speech4	0.50	0.85	**0.85**	**0.97**	0.80	0.96
Speech8	0.46	0.83	**0.83**	**0.96**	0.74	0.94
Speech10	0.46	0.83	**0.81**	**0.96**	0.72	0.94
Speech20	0.48	0.83	**0.60**	**0.90**	0.52	0.87
10halo	0.53	0.86	**0.85**	**0.96**	0.82	0.95
20depspeech	0.53	0.86	**0.85**	**0.96**	0.81	0.95
nband5	0.46	0.82	**0.86**	0.96	0.86	**0.96**
Gnband	0.44	0.82	**0.45**	**0.82**	0.45	0.82
EEG19	0.47	0.83	0.90	**0.97**	**0.90**	0.97

Fig. 8. The 95% efficiency level and sumeff for the ICALAB benchmarks.

Thus, as we require the W-DO assumption, and we wish to select the best window for the sources in terms of W-DO given the mixture, we measure the efficiency of the mixture representations and select the window size that is most efficient as we hope, based on the experimental results, that in that representation the sources will be maximally W-DO.

Appendix B - Algorithm Initialization

One way to initialize the mixing matrix in Section 4 instead of using random values is to cluster a small random selection of time-frequency points from those which make up the 95% efficiency. For each of these points, we consider the D dimensional vector $X(k,\omega)$. Under the W-DO assumption, the $X(k,\omega)$ should be γA_l for some $\gamma \in \mathbf{C}$ for some $l \in \{1,\ldots,L\}$. We construct a distance matrix between all pairs of $X(k,\omega)$ using the following metric. The distance between vector $\mathbf{a}$ and $\mathbf{b}$ is,

$$d(\mathbf{a},\mathbf{b}) = \sqrt{\|\mathbf{a}\|^2\|\mathbf{b}\|^2 - |\mathbf{a}\cdot\mathbf{b}|} \tag{27}$$

which has the important property that $d(\mathbf{a},\mathbf{b}) = 0$ iff $\mathbf{b} = \alpha\mathbf{a}$, for some $alpha \in \mathbf{C}$. That is, pairs of observations which lie on a line through the origin are consistent with the W-DO assumption. Clusters are formed using the pairwise distance matrix and MATLAB's "cluster" function [14]. We cluster on a small random selection of points instead of all the points because of time constraints. In practice, we randomly select 300 points from those time-frequency points making up the 95% efficiency (instead of simply selecting the largest 300 time-frequency components) because often the largest components will be dominated by a subset of the sources. Considering points making up the 95% efficiency helps to ensure that even the lower power sources have representation in the initialization.

7. REFERENCES

[1] A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. Wiley, April 2002.

[2] A. Belouchrani and M. Amin. Blind source separation based on time-frequency signal representations. *IEEE Trans. on Signal Processing*, 46(11):2888–2897, November 1998.

[3] A. Cichocki, S. Amari, and K. Siwek et al. *ICALAB Toolboxes,* `http://www.bsp.brain.riken.go.jp/ICALAB`.

[4] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. on Signal Processing*. Submitted November 2002.

[5] S. Rickard, R. Balan, and J. Rosca. Real-time time-frequency based blind source separation. In *Proc. ICA*, pages 651–656, 2001.

[6] S. Rickard and O. Yilmaz. On the W-disjoint orthogonality of speech. In *Proc. ICASSP*, volume 1, pages 529–532, 2002.

[7] P. Bofill and M. Zibulevsky. Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform. In *Proc ICA*, pages 87–92, Helsinki, Finland, June 19–22 2000.

[8] M. Aoki, M. Okamoto, S. Aoki, and H. Matsui. Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones. *Acoust. Sci. & Tech.*, 22(2):149–157, 2001.

[9] L.-T. Nguyen, A. Belouchrani, K. Abed-Meraim, and B. Boashash. Separating more sources than sensors using time-frequency distributions. In *Int. Symp. on Sig. Proc. and its Applications (ISSPA)*, pages 583–586, Kuala Lumpur, Malaysia, August 13–16 2001.

[10] S. T. Roweis. One microphone source separation. In *Neural Information Processing Systems 13 (NIPS)*, pages 793–799, 2000.

[11] B. Berdugo, J. Rosenhouse, and H. Azhari. Speakers' direction finding using estimated time delays in the frequency domain. *Signal Processing*, 82:19–30, 2002.

[12] P. Bofill. Underdetermined blind separation of delayed sound sources in the frequency domain. *preprint*, 2002.

[13] R. Balan, J. Rosca, and S. Rickard. Scalable non-square blind source separation in the presence of noise. In *sent to ICASSP2003, Hong-Kong, China*, April 2003.

[14] *MATLAB Statistics Toolbox,* `http://www.mathworks.com`.