

Survey of Sparse and Non-Sparse Methods in Source Separation

Paul D. O'Grady,¹ Barak A. Pearlmutter,¹ Scott T. Rickard²

¹ Hamilton Institute, National University of Ireland Maynooth, Co. Kildare, Ireland. Ollscoil na hÉireann Má Nuad, Éire

² University College Dublin, Ireland. An Coláiste Ollscoile, Baile Átha Cliath, Éire

Received 20 September 2004; accepted 4 April 2005

ABSTRACT: Source separation arises in a variety of signal processing applications, ranging from speech processing to medical image analysis. The separation of a superposition of multiple signals is accomplished by taking into account the structure of the mixing process and by making assumptions about the sources. When the information about the mixing process and sources is limited, the problem is called 'blind'. By assuming that the sources can be represented sparsely in a given basis, recent research has demonstrated that solutions to previously problematic blind source separation problems can be obtained. In some cases, solutions are possible to problems intractable by previous non-sparse methods. Indeed, sparse methods provide a powerful approach to the separation of linear mixtures of independent data. This paper surveys the recent arrival of sparse blind source separation methods and the previously existing non-sparse methods, providing insights and appropriate hooks into the literature along the way. © 2005 Wiley Periodicals, Inc. *Int J Imaging Syst Technol*, 15;18–33;2005; Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/ima.20035

Key words: Blind Sources Separation; sparse methods; Non-negative Matrix Factorization

I. INTRODUCTION

When presented with a set of observations from sensors such as microphones, the process of extracting the underlying sources is called *source separation*. Doing so without strong additional information about the individual sources or constraints on the mixing process is called *blind source separation* (BSS). Generally the problem is stated as follows: given M linear mixtures of N sources mixed via an unknown $M \times N$ mixing matrix \mathbf{A} , estimate the underlying sources from the mixtures. When $M \geq N$, this can be achieved by constructing an unmixing matrix \mathbf{W} , where $\mathbf{W} = \mathbf{A}^{-1}$ up to permutation and scaling of the rows. The conditions that must be satisfied to guarantee separation are given by Darmois' Theorem (Darmois, 1953), which states that the sources must be non-Gaussian and statistically independent. The dimensionality of the mixing process

influences the complexity of source separation. If $M = N$, the mixing process \mathbf{A} is defined by an even-determined (i.e. square) matrix and, provided that it is non-singular, the underlying sources can be estimated by a linear transformation. If $M > N$, the mixing process \mathbf{A} is defined by an over-determined matrix and, provided that it is full rank, the underlying sources can be estimated by least-squares optimisation or linear transformation involving matrix pseudo-inversion. If $M < N$, then the mixing process is defined by an under-determined matrix and consequently source estimation becomes more involved and is usually achieved by some non-linear technique.

Environmental assumptions about the surroundings in which the sensor observations are made also influence the complexity of the problem. BSS of acoustic signals is often referred to as the *cocktail party problem* (Cherry, 1953); that is the separation of individual voices from a myriad of voices in an uncontrolled acoustic environment such as a cocktail party. Sensor observations in a natural environment are confounded by signal reverberations, and consequently, the estimated unmixing process needs to identify a source arriving from multiple directions at different times as one individual source. Generally, BSS techniques depart from this difficult real-world scenario and make less realistic assumptions about the environment so as to make the problem more tractable. There are typically three assumptions that are made about environment. The most rudimentary of these is the *instantaneous case*, where sources arrive instantly at the sensors but with differing signal intensity. An extension of the previous assumption, where arrival delays between sensors are also considered, is known as the *anechoic case*. The anechoic case can be further extended by considering multiple paths between each source and each sensor, which results in the *echoic case*, sometimes known as *convolutional mixing*. Each case can be extended to incorporate linear additive noise, which is usually assumed to be white and Gaussian.

Assumptions can also be made about the nature of the sources. Such assumptions form the basis for most source separation algorithms and include statistical properties such as independence and stationarity. One increasingly popular and powerful assumption is that the sources have a parsimonious representation in a given basis.

Correspondence to: Paul D. O'Grady; e-mail: paul.ogrady@nuim.ie
Supported by Higher Education Authority of Ireland (An túdarás Um Ard-Oideachas) and Science Foundation Ireland grant 00/PI.1/C067.

Table I. A Comparison of Various Source Separation Techniques.

Algorithm	No. of Sources (M)			No. of Sensors (M)			Mixing Model			Representation			Unmixing			
	M	$>M$	\times	1	2	>2	Inst.	Anechoic	Echoic	Time	Freq.	Signal Dict.	Parameter Est.	Separation		
														Linear filtering	Adaptive feedback network	Linear transformation
Wiener (1949) Filter	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times		Power spectrum reshaping	Linear filtering		
Herault and Jutten (1986)	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times		Independence maximisation	Adaptive feedback network		
JADE (Cardoso and Souloumiac, 1993)	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times		Joint cumulant diagonalisation	Linear transformation		
ICA (Comon, 1994)	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times		Mutual information minimisation	Linear transformation		
BS-InfoMax (Bell and Sejnowski, 1995)	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times		Entropy maximisation	Linear transformation		
Lambert (1995)	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times		Multichannel blind deconvolution	Current estimate feedback		
Lin et al. (1997) ^a	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times		Hough transform	Hard assignment		
SOBI (Belouchrani et al., 1997)	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times		Joint unitary diagonalisation	Linear transformation		
fICA (Hyvärinen and Oja, 1997)	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times		Non-Gaussianity maximisation	Linear transformation		
Lee et al. (1999) ^a	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times		Gradient ascent learning	Maximum a posteriori		
DUJET (Jourjine et al., 2000) ^a	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times		2D Histogram clustering	Binary time-frequency masking		
Bofill and Zibulevsky (2000) ^a	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times		Potential function clustering	L_1 -Norm minimisation		
Zibulevsky and Pearlmutter (2001) ^a	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times		MAP with Laplacian prior	Joint optimisation		
Roweis (2001)	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times		Hidden markov models	Spectral masking and filtering		
Pearlmutter and Zador (2004) ^a	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times	\times		Known HRTF and dictionary	L_1 -Norm minimisation		

^aSparse methods.

These methods have come to be known as *sparse methods*. A signal is said to be sparse when it is zero or nearly zero more than might be expected from its variance. Such a signal has a probability density function or distribution of values with a sharp peak at zero and fat tails. This shape can be contrasted with a Gaussian distribution, which would have a smaller peak and tails that taper quite rapidly. A standard sparse distribution is the Laplacian distribution ($p(c) \propto e^{-|c|}$), which has led to the sparseness assumption being sometimes referred to as a Laplacian prior. The advantage of a sparse signal representation is that the probability of two or more sources being simultaneously active is low. Thus, sparse representations lend themselves to good separability because most of the energy in a basis coefficient at any time instant belongs to a single source. This statistical property of the sources results in a nicely defined structure being imposed by the mixing process on the resultant mixtures, which can be exploited to make estimating the mixing process much easier. Additionally, sparsity can be used in many instances to perform source separation in the case when there are more sources than sensors. A sparse representation of an acoustic signal can often be achieved by a transformation into a Fourier, Gabor or Wavelet basis.

The sparse representation of an acoustic signal has an interpretation in information theoretic terms, where the representation of a signal using a small number of coefficients corresponds to transmission of information using a code utilising a small number of bits. Sparse representation of information is a phenomenon that also occurs in the natural world. In the brain, neurons are said to encode data in a sparse way, if their firing pattern is characterised by long periods of inactivity (Földiák and Young, 1995; Körding et al., 2002), and recent work (DeWeese et al., 2003) indicates that sparse representations exist in the auditory cortex.

Blind source separation has been studied for nearly two decades. The earliest approach traces back to Herault and Jutten (1986) whose goal was to separate an instantaneous linear even-determined mixture of non-Gaussian independent sources. They proposed a solution that used a recurrent artificial neural network to separate the unknown sources, and the crucial assumption being that the underlying signals were independent. This early work led to the pioneering adaptive algorithm of Jutten and Herault (1991). Linsker (1989) proposed unsupervised learning rules based on information theory that maximise the average mutual information between the inputs and outputs of an artificial neural network. Comon (1994) proposed that mutual information was the most natural measure of independence and showed that maximising the non-Gaussianity of the source signals was equivalent to minimising the mutual information between them. He also dubbed the concept of determining underlying sources by maximising independence, *independent component analysis* (ICA). Bell and Sejnowski (1995) developed a BSS algorithm called *BS-Infomax*, which is similar in spirit to that of Linsker and uses an elegant stochastic gradient learning rule that was proposed by Amari et al. (1996). The idea of non-Gaussianity of sources was used by Hyvärinen and Oja (1997) to develop their *fICA* algorithm. As an alternative approach to separation using mutual information, Gaeta and Lacoume (1990) proposed maximum likelihood estimation, an approach elaborated by Pham et al. (1992), although, Pearlmutter and Parra (1996) and Cardoso (1997) later demonstrated that the BS-Infomax algorithm and maximum likelihood estimation are essentially equivalent. The early years of BSS research concentrated on solutions for even-determined and over-determined mixing processes. It was not until recent years that a solution for the under-determined case was proposed when Belou-

chrani and Cardoso (1994) presented a *maximum a posteriori* (MAP) probability approach for discrete QAM sources. An approach for sparse sources was later proposed by Lewicki and Sejnowski (1998). The first practical algorithm for separation in an anechoic environment was the *DUET* algorithm initially proposed by Jourjine et al. (2000) and further explored by Yilmaz and Rickard (2004). The first algorithms for the anechoic separation of moving speakers were presented by Rickard et al. (2001) and Anemüller and Kollmeier (2003). A selection of sparse and non-sparse source separation algorithms and their characteristics are presented in Table I.

Blind source separation techniques are not confined to acoustic signals. BSS has also been applied to the decomposition of functional brain imaging data such as electroencephalography (Jung et al., 1999, 2000), functional magnetic resonance imaging (McKeown et al., 1998), and magnetoencephalography (Tang et al., 2000; Vigário et al., 2000; Wübbeler et al., 2000; Ziehe et al., 2000; Pearlmutter and Jaramillo, 2003). BSS has also been applied to such diverse areas as real time robot audition (Nakadai et al., 2002), digital watermark attacks (Du et al., 2002), and financial time series analysis (Roth and Baram, 1996; Back and Weigend, 1997). It has even been conjectured that BSS will have a role in the analysis of the Cosmic Microwave Background (Cardoso et al., 2003), potentially helping to elucidate the very origins of the universe.

This paper presents an overview of BSS techniques and is organised as follows. The mixing models specific to each of the three cases specified earlier are presented in Section II and then a staged algorithm approach is discussed; staged algorithms perform mixing parameter estimation and source separation independently. Because these stages are separate, any method from stage one can be combined with any method in stage two to form a source separation algorithm. Thus, in Section III, we present a selection of mixing parameter estimation techniques, and in Section IV, we present the unmixing methods for these staged algorithms. Within these sections, we divide the techniques presented into instantaneous, anechoic and echoic mixing model subsections. Alternatively, some source separation techniques estimate mixing parameters and sources simultaneously; these methods are presented in Section V.

II. Mixing Models The generative model for the BSS problem can be presented as follows: a set of T observations of M sensors

$$\mathbf{X} = [\mathbf{x}(1)| \cdots | \mathbf{x}(T)] = \begin{bmatrix} x_1(0) & x_1(p) & \cdots & x_1((T-1)p) \\ x_2(0) & x_2(p) & \cdots & x_2((T-1)p) \\ \vdots & \vdots & \ddots & \vdots \\ x_M(0) & x_M(p) & \cdots & x_M((T-1)p) \end{bmatrix}$$

consist of a linear mixture of N source signals

$$\mathbf{S} = [\mathbf{s}(1)| \cdots | \mathbf{s}(T)] = \begin{bmatrix} s_1(0) & s_1(p) & \cdots & s_1((T-1)p) \\ s_2(0) & s_2(p) & \cdots & s_2((T-1)p) \\ \vdots & \vdots & \ddots & \vdots \\ s_N(0) & s_N(p) & \cdots & s_N((T-1)p) \end{bmatrix}$$

by way of an unknown linear mixing process characterised by the $M \times N$ mixing matrix \mathbf{A}

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MN} \end{bmatrix}$$

yielding the equation

$$\mathbf{x}(t) = \mathbf{A} \star \mathbf{s}(t) + \epsilon(t), \quad (1)$$

where $\epsilon(t)$ is noise (usually white and Gaussian), p is the sample period, and \star denotes the model dependent linear operator. The form of the elements of the mixing matrix, a_{ij} , and the linear operator in the above equation are mixing model dependent and define whether the mixing process is instantaneous, anechoic or echoic. Table II presents the linear operators and mixing matrix elements specific to the three cases of BSS, where the operator $\delta(t - \delta_{ij})$ is used to denote a delay between source j to sensor i , c_{ij} is a scalar attenuation factor between source j to sensor i , δ_{ij}^k and c_{ij}^k are the delay and attenuation parameters for the k -th arrival path, and L is the number of paths the sources can take to the sensors.

The following subsections detail the three cases of BSS. In a slight abuse of conventional notation, the temporal index t is used for the relevant domain, which will be made explicit only when necessary.

A. Instantaneous Mixing. For the instantaneous case, the mixing matrix \mathbf{A} simply consists of scalars representing signal amplitudes. Taking a simple example where there are two sources and two mixtures, the generative model takes the form

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \end{bmatrix}, \quad (2)$$

and can be described as a linear mixture of N linear subspaces in M -space. This linear mixing imposes a structure on the resultant mixtures, which becomes apparent when the mixtures have a sparse representation (see Fig. 1). The existence of this structure can be explained as follows: from Eq. (2) it is evident that if only one source is active, say s_1 , then the resultant mixtures would be

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = s_1(t) \begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix}.$$

Therefore, the points on the scatter plot of $x_1(t)$ versus $x_2(t)$ would lie on the line through the origin whose direction is given by the vector $[a_{11} \ a_{21}]^T$. When the sources are sparse, making it unusual for more than one source to be active at the same time, the scatter plot of coefficients would constitute a mixture of lines, with the lines broadened because of noise and the occasional simultaneous activity. These line orientations are unique to each source and correspond to the columns of the mixing matrix \mathbf{A} ; therefore, the essence of the sparse approach is the identification of line orienta-

Table II. Mixing Model Specific Linear Operators and Mixing Parameters.

Mixing Model	Linear operator	Generative Model	a_{ij}
Instantaneous	Matrix multiply	$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$	c_{ij}
Anechoic	Convolution	$\mathbf{x}(t) = \mathbf{A} \star \mathbf{s}(t)$	$c_{ij} \delta(t - \delta_{ij})$
Echoic	Convolution	$\mathbf{x}(t) = \mathbf{A} \star \mathbf{s}(t)$	$\sum_{k=1}^L c_{ij}^k \delta(t - \delta_{ij}^k)$

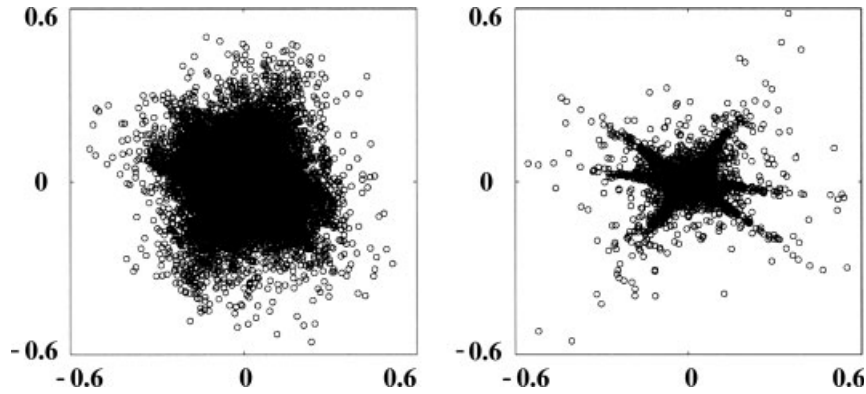


Figure 1. Scatter plot of two linear mixtures of three zero-mean sources in both the time domain (left) and the transform domain (right). The ‘sparse’ transform domain consists of the coefficients of 512-point windowed FFTs. The figures axes are measured in arbitrary units of transform domain mixture coefficients.

tions vectors (also known as basis vectors) from the observed data. In contrast, traditional non-sparse approaches exploit the statistics of the sources as opposed to the structure of the mixtures and employ methods such as mutual information minimisation and independence maximisation to identify the mixing process (see Table I).

B. Anechoic Mixing. Anechoic mixing is an extension of instantaneous mixing in which source transmission delays between sensors are also considered. Taking a simple example with two mixtures and N sources, the generative model is

$$x_1(t) = \sum_{j=1}^N a_{1j} s_j(t - \delta_{1j}), \quad (3)$$

$$x_2(t) = \sum_{j=1}^N a_{2j} s_j(t - \delta_{2j}), \quad (4)$$

where the attenuation and delay of source j to sensor i would be determined by the physical position of the source relative to the sensors. Some algorithms make an assumption that the distance between the sensors is smaller than half the wavelength of the highest frequency of interest. This assumption, referred to in the array signal processing literature as the narrow band assumption, makes

it possible for a delay between sensors to be represented unambiguously as a phase shift of a signal. The problem of anechoic signal unmixing is therefore to identify the attenuation factor and relative delay associated with each source. An illustration of the anechoic case is provided in Figure 2.

C. Echoic Mixing. The echoic case of BSS considers not only transmission delays but reverberations too. This results in a more involved generative model that in turn makes finding a solution more difficult. A parametric mixing model with two sources and two mixtures takes the form

$$x_1(t) = \sum_{k=1}^L a_{11}^k s_1(t - \delta_{11}^k) + a_{12}^k s_2(t - \delta_{12}^k), \quad (5)$$

$$x_2(t) = \sum_{k=1}^L a_{21}^k s_1(t - \delta_{21}^k) + a_{22}^k s_2(t - \delta_{22}^k), \quad (6)$$

where L is the number of paths the source signal can take to the sensors. The mixing matrix \mathbf{A} can be thought of as a matrix of finite impulse response (FIR) filters that are defined by the environment

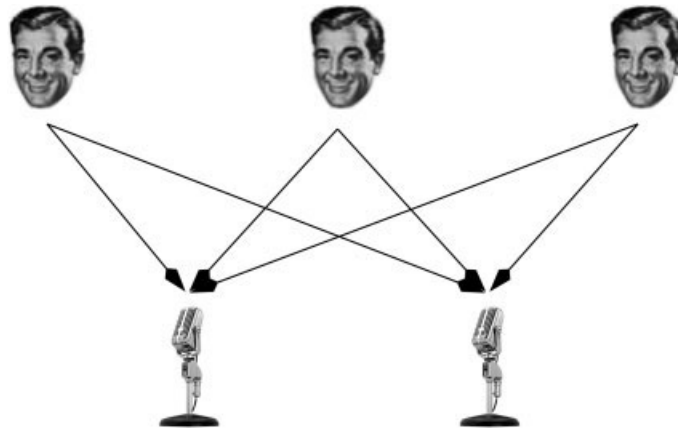


Figure 2. Anechoic mixing: speech is observed at the microphones with differing intensity and arrival times (because of propagation delays) but with no reverberations.

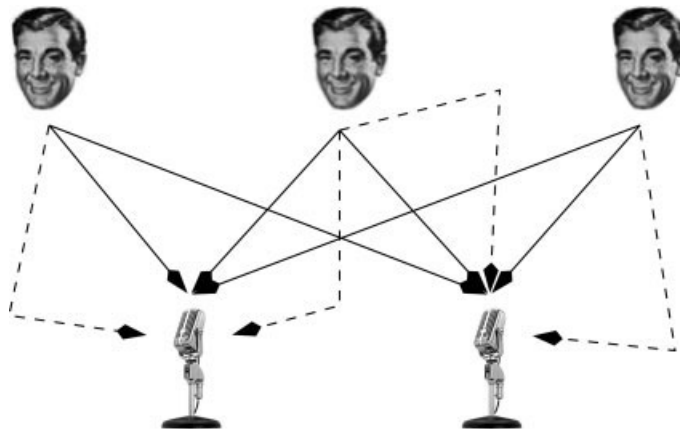


Figure 3. Echoic mixing: speech is observed at the microphones with differing arrival times, intensity and multiple arrival paths because of reverberations.

in which the observations are made, and observations are generated by convolving the sources with this FIR matrix

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix} * \begin{bmatrix} s_1(t) \\ s_2(t) \end{bmatrix}, \quad (7)$$

where \tilde{A} represents an individual FIR filter. Therefore, the problem of unmixing is to find the appropriate filters for the model so that the source signals can be unmixed and deconvolved.

An illustration of the echoic case is provided in Figure 3.

III. MIXING PARAMETERS ESTIMATION

In a staged algorithm approach, the first step is to estimate \mathbf{A} , the form of which is dependent on the environment considerations and the dimensionality of the problem. The following subsections explore the complexity of estimating \mathbf{A} using sparse and non-sparse methods in the context of instantaneous, anechoic and echoic mixing assumptions.

A. Instantaneous Mixing

A.1. Sparse Methods. The following methods illustrate how the underlying mixing structure can be used to estimate the mixing matrix using the sparseness assumption. These methods estimate the mixing process in a piecemeal fashion, where each column of the mixing matrix is determined individually; thus, enabling a solution to be found for the under-determined case. Each method involves a transformation of the observed data into a sparse basis using either the Fourier or Wavelet transform.

One approach that can be used to identify line in a scatter plot is to use a clustering technique. Zibulevsky et al. (2002) perform clustering by normalising all observed data points and then mapping them to the unit hemisphere, this mapping being required because the line orientation for each source exists in both hemispheres, producing two clusters for each source. Each cluster pair is consolidated by this mapping and line orientations are represented as clusters of data points on the unit hemisphere.* A fuzzy C-Means algo-

*Except when line orientation lies on the equator in which case the mapping can fail to consolidate its two halves.

rithm is used to identify the cluster centres, which are adjoined to create the estimated mixing matrix $\hat{\mathbf{A}}$.

Another clustering based approach in which topographic maps are used is presented by van Hulle (1999). Two data observations are represented in a polar coordinate form, which results in line orientations creating clusters. An unsupervised learning technique called a topographical map, which uses a kernel-based maximum entropy learning rule (Hulle, 1998), is used to learn the features of the mixture distribution. When the topographic map has converged, the density estimate of the mixture distribution is calculated from the weights of the map and is used to estimate the mixing matrix. This technique was shown to unmix time domain instantaneous mixtures of speech, demonstrating that speech is sparse in the time domain. However, speech is more sparse in the time-frequency domain (Rickard and Yilmaz, 2002).

The line orientation of a data set can be thought of as the direction of its greatest variance. One way to find this direction is to perform an eigenvector decomposition on the covariance matrix of the data, the resultant principal eigenvector i.e. the eigenvector with the largest eigenvalue, indicates the direction of the data. This calculation forms the basis of the *principal component analysis* (PCA) (Pearson, 1901) dimensionality reduction technique, which is based on second order statistics, and is considered here in a degenerate form in which only one principal component is retained. In the case of identifying lines in a scatter plot where the line orientations are not orthogonal, N different line orientations exist, which require separation of the data into N disjoint sets, each with an individual PCA computation. It is also worth noting that PCA is used in the pre-whitening step of many algorithms and that methods exist that are equivalent to PCA for fourth order statistics that achieve separation of sources; the problem here is the diagonalisation of a quadricovariance tensor (Cardoso, 1990).

A modified *k-Means* (MacQueen, 1967) algorithm can be used to identify lines in a scatter plot (O'Grady and Pearlmutter, 2004a). The approach can be described as follows: the algorithm is randomly initialised with N line orientation estimates \mathbf{v}_i and each data point is hard assigned to the closest line orientation estimate. Each line orientation has a stochastic gradient algorithm associated with it, which is updated using the new data point assigned to it. The stochastic gradient algorithm is an online PCA technique, which calculates an estimate of the principal eigenvector of the data, which is then used as the new line orientation estimate. The procedure is



Figure 4. Separation of 6 sources from 4 mixtures using the Soft-LOST algorithm (O’Grady and Pearlmutter, 2004b). The plots above show 10-s clips of six acoustic sources s_1, \dots, s_6 , 4 mixtures x_1, \dots, x_4 and 6 source estimates $\hat{s}_1, \dots, \hat{s}_6$, where sound wave pressure is plotted against time in units of seconds. The signal-to-noise ratios of the estimated sources in dB are as follows: 10.5845, 13.3895, 13.7402, 8.9102, 15.2732 and 11.3213.

repeated over all data points where each stochastic gradient algorithm converges to an individual line orientation, which are then adjoined to form the estimated mixing matrix, $\hat{A} = [v_1 | \dots | v_N]$. This algorithm is analogous to k -Means where line orientations and distances from a line replace cluster centres and distance from the cluster centres of conventional k -Means. This idea can be modified to include soft data assignments and a batch mode of operation, where the stochastic gradient algorithm is replaced by an eigenvector decomposition of the covariance matrix of the data set (O’Grady and Pearlmutter, 2004b). This approach is an *expectation maximisation* (EM) algorithm (Dempster et al., 1976), where the E-step calculates posterior probabilities assigning data points to lines and M-step repositions the lines to match the points assigned to them. The separation of 6 sources from 4 mixtures using this EM based approach is presented in Figure 4. Theis (2001) presents a similar approach and Kearns et al. (1997) contains a useful discussion of hard and soft assignments.

Feature detection techniques from image processing have also been used to locate line orientations in a scatter plot. Lin et al. (1997) present an algorithm that uses a Hough transform to identify lines. The scatter plot data are partitioned into bins, providing an

image representation, the image is then convolved with an edge detection operator, and the resultant image is normalised and thresholded to form a binary image. This image is then Hough transformed using line orientations as the feature of interest and the line orientation vectors of the scatter plot are identified as peaks in the Hough transform space. The line orientation parameters associated with each peak in Hough space are then adjoined to form the estimated mixing matrix.

Bofill and Zibulevsky (2001) present a method that is restricted to two observations. A radial grid of K lines that covers the space of all possible directions (i.e. from 0 to π radians) is imposed on the scatter plot, the difference between the polar phase of all data points and one radial grid line is calculated, and each value is used to parameterise a local kernel density function (sometimes called a potential function). Each of the T local kernel estimates is then combined to give a single kernel estimate for that grid line. Grid lines that are close to scatter plot line orientations are highly activated compared with those that are not. The procedure is repeated for all grid lines and the result is a global kernel density function spanning K -grid lines where the local maxima of the function correspond to the columns of the mixing matrix.

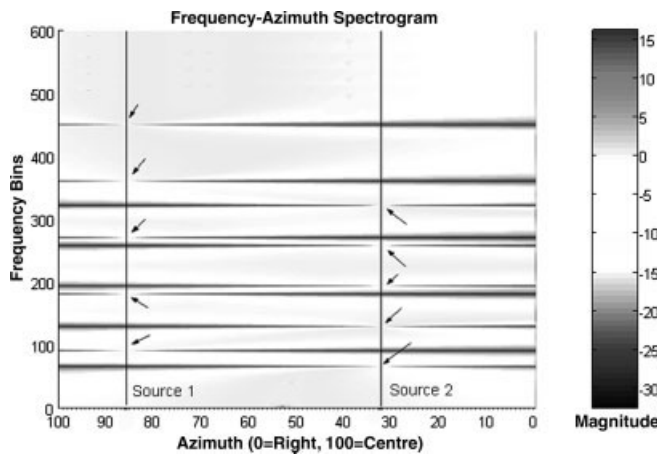


Figure 5. The frequency-azimuth spectrogram for the right channel. Two synthetic sources comprising 5 non-overlapping partials are used. The arrows indicate frequency dependent nulls caused by phase cancellation. (From Barry et al., 2004, Figure 1, with permission).

Source separation techniques specific to the separation of sources within a stereophonic music recording also exist. Barry et al. (2004) present an algorithm that exploits the fact that stereophonic music recordings are mixed using a panoramic potentiometer, which positions individual sources (instruments) within the stereo field. Gain-scaling is applied to one mixture so that one source's intensity becomes equal in both the left and right mixture. Subtraction of the mixtures will cause the source to cancel out because of phase cancellation. The appropriate gain required for the gain-scaling operation is determined by creating a frequency-azimuth spectrogram; both mixtures are transformed using a discrete Fourier transform (DFT) of a specified frame size and the left mixture is scaled by a range of values corresponding to 0° (far left) to 180° (far right) azimuth. The resultant frames are subtracted from the right mixture frame and the results are used to construct the frequency-azimuth spectrogram. This procedure is repeated for the left mixture. As can be seen from Figure 5, nulls or local minima are formed in the frequency bands specific to each source. These nulls represent the gain needed to cancel each source. In this way, the position (mixing parameters) of each source can be determined by visual inspection or by using some optimisation technique.

A.2. Non-Sparse Methods. The following methods do not make the sparseness assumption and instead estimate the mixing process by exploiting the statistics of the observations in the time domain. These methods are restricted to the even-determined case.

The BS-Infomax algorithm of Bell and Sejnowski (1995) shows that for signals with a positive kurtosis, such as speech, minimising the mutual information between the source estimates and maximising the entropy of the source estimates are equivalent. Entropy maximisation can then be implemented using a stochastic gradient ascent rule. There are a couple of interesting aspects to the approach. First, the model is parameterised not by the mixing matrix, as is natural for a forward model, but rather by its inverse i.e. the unmixing matrix. Although the inverse contains the same information, it makes for a much simpler update rule and superior numeric conditioning. The second interesting aspect is that the algorithm takes a naive stochastic gradient descent algorithm and modi-

fies it by multiplying the gradient by a particular positive-definite matrix, $\mathbf{W}^T\mathbf{W}$, where \mathbf{W} is the current estimate of the unmixing matrix. While Amari et al. (1996) provide a theoretical justification of this based on information geometry, a naive perspective would be that the multiplier chosen fortuitously eliminates a matrix inversion, making each step of the algorithm much faster and the convergence rate independent of the condition number of the unmixing matrix.

Parra and Sajda (2003) formalise the problem of finding the unmixing matrix as a generalised eigenvector decomposition of two matrices, where one of the matrices is the covariance matrix of the observations and the other is a symmetric matrix whose form depends on the statistical assumptions on the sources (non-stationary decorrelated, non-white decorrelated or non-Gaussian independent). This formulation combines subspace analysis and mixing matrix construction into a single computation, making for simple implementation.

Algorithms that rely solely on second order statistics have also been developed, one such algorithm is the *SOBI* algorithm (Belouchrani et al., 1997). *SOBI* is based on the unitary diagonalisation of the whitened data covariance matrix. This idea is extended to exploit the time coherence of the original sources, where multiple whitened covariance matrices are diagonalised simultaneously, and each matrix is calculated from observation data taken at different time delays. The algorithm considers the effect of noise on the observations by using a whitening matrix calculation that improves robustness to noise. Unitary diagonalisation can be explained as follows: given a whitening matrix \mathbf{V} and the observations \mathbf{X} , the covariance matrix of the whitened observations is $E[\mathbf{V}\mathbf{X}\mathbf{X}^H\mathbf{V}^H] = \mathbf{V}\Sigma_{\mathbf{X}}\mathbf{V}^H = \mathbf{V}\mathbf{A}\Sigma_{\mathbf{S}}\mathbf{A}^H\mathbf{V}^H = \mathbf{I}$, where superscript \mathbf{H} denotes the complex conjugate transpose of a matrix, $\Sigma_{\mathbf{X}}$ denotes the observed data covariance matrix, and $\Sigma_{\mathbf{S}}$ denotes the covariance matrix of the original sources. The source signals are assumed to have unit variance and are uncorrelated, therefore $\Sigma_{\mathbf{S}} = \mathbf{I}$. It turns out that $\mathbf{V}\mathbf{A}$ is a unitary matrix, therefore $\mathbf{V}\mathbf{A} = \mathbf{U}$ and the mixing process can be factored as $\mathbf{A} = \mathbf{V}^{-1}\mathbf{U}$. The problem is then to find a unitary matrix that diagonalises the whitened covariance matrix, and for the case where there are multiple covariance matrices, the problem is to find a unitary matrix that simultaneously diagonalises all matrices generated from the different time delays. This joint approximate diagonalisation operation can be computed efficiently using a generalisation of the Jacobi technique for the exact diagonalisation of a single Hermitian matrix.

Sometimes a distinction is made between BSS and ICA techniques; for the purposes of the paper, we consider ICA to be a restrictive case of BSS where only the even-determined or over-determined case is considered, the mixing process is instantaneous and the observed data is pre-whitened. Consequently the mixing matrix \mathbf{A} is factorised into $\mathbf{A} = \mathbf{V}\mathbf{R}$, where \mathbf{V} is the whitening matrix and \mathbf{R} is an orthogonal rotation. The factor \mathbf{R} is determined by assuming independence between the original sources and selecting a matrix that will maximise independence in the observations, using some independence criteria. The *fICA* algorithm of Hyvärinen and Oja (1997) uses an independence criteria that forces non-Gaussianity of the estimated sources. Other criteria such as maximum likelihood (Gaeta and Lacoume, 1990) and mutual information (Comon, 1994) are presented in the literature; for an in depth discussion on techniques particular to our definition of ICA, see Hyvärinen (1999) and Cardoso (1998).

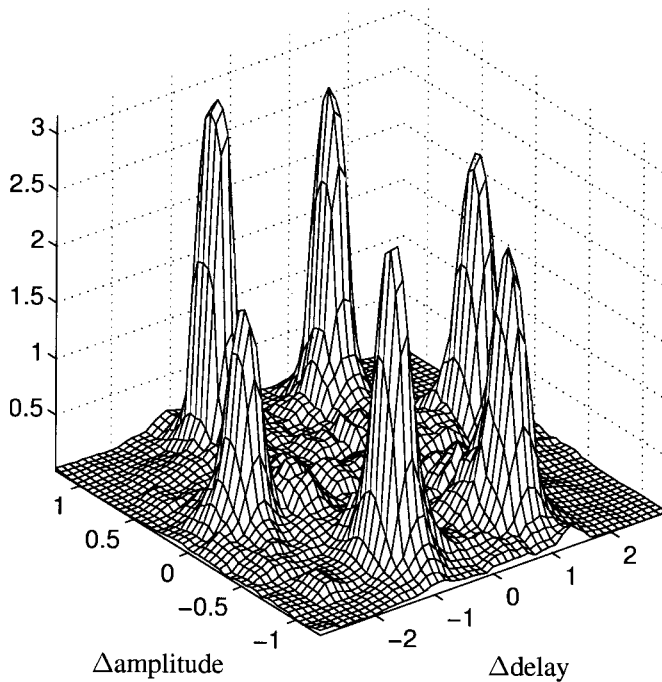


Figure 6. The DUET smoothed power-weighted histogram for a six speech source mixture. Each peak corresponds to one source and the peak locations correspond to the associated source's mixing parameters. (From Yilmaz and Rickard, 2004, Figure 8, with permission).

B. Anechoic Mixing

B.1. Sparse Methods. A simple online approach to anechoic unmixing is presented by Aoki et al. (2001). Two sensor observations are transformed into the Fourier domain, and the amplitude and phase differences between the frequency bins of each observation are calculated. The amplitude and phase associated with each source is known in advance, and if the difference between these ideal values and the current calculated values is smaller than some specified threshold, then the associated frequency bin coefficient is assigned to that source. Although this method could not be described as a blind model, it gives a good insight into how sources can be separated in an anechoic environment.

A more sophisticated blind approach called the DUET algorithm is presented in Jourjine et al. (2000) and further explored in Yilmaz and Rickard (2004). The observations of each sensor are transformed into the time–frequency domain using a short-time windowed Fourier transform, and the relative attenuation and delay values between the two observations are calculated from the ratio of corresponding time–frequency points. The main assumption is that, because of the sparsity of speech in the time–frequency domain, almost all mixture time–frequency points with significant magnitude are in fact due to only one of the original sources. Thus, each point in the stereo mixing is essentially a mixture of at most one source, and the goal is to calculate to which source each time–frequency point belongs. To do this, the relative attenuation and delay from each ratio of time–frequency points are placed into a 2D power-weighted histogram. Each source will cause a peak in the histogram. These peaks correspond to the relative attenuation and delay mixing parameters of the sources. An attenuation–delay histogram that indicates the existence of six sources and their associated attenuation and

delay parameters can be seen in Figure 6. Rickard et al. (2001) present an online algorithm based on DUET that uses a maximum likelihood-based gradient search method to track the mixing parameters. Choi (2003) presents a real-time approach to anechoic unmixing using a method similar to DUET, where the 2D histogram of estimates is replaced by a k -Means clustering algorithm.

Scatter plot representations can also be used in the anechoic case. Bofill (2002) presents a method for anechoic unmixing that uses the same line orientation identification approach used in the instantaneous case. The generative model is modified slightly in which the attenuation and delay parameters of the mixing process are segregated into two matrices. Observed data is transformed using a DFT and the magnitude of the resultant complex data is used to create a scatter plot. The line orientations of the scatter plot are identified by the kernel density estimation method of Bofill and Zibulevsky (2001) (discussed in Section IIIA) and the resultant line orientation vectors are adjoined to form the attenuation matrix. The estimated delay matrix is formed by taking the real and imaginary coefficients assigned to source j in the previous operation and iteratively rectifying the delay parameter until the kernel function of the data is maximised. The procedure is repeated for the N sources and the resultant delay parameters form the estimated delay matrix.

B.2. Non-Sparse Methods. A number of modifications that allow existing instantaneous case algorithms to be used for anechoic situations have been proposed. Platt and Faggin (1992) apply the mathematics of adaptive noise cancelling networks (Widrow et al., 1975) to the source separation network of Héault and Jutten (1986). The adaption of noise cancellation networks is based on an elegant notion: if a signal is corrupted by noise, it will have higher power than when it is not, because the noise power adds to the signal power. This principle known as the *minimum output power* principle is used to adapt an extension of the Héault–Jutten network that contains delays in the feedback path. These delays cancel out interference from the delayed versions of other sources in the mixtures and the new network is then capable of unmixing delayed mixtures. This model can also be extended to the echoic case.

The BS-Infomax algorithm has also been extended to the anechoic case. Torkkola (1996b) extends BS-Infomax using the adaptive delay architecture of Platt and Faggin (1992). The resultant network maximises the entropy at the output of the network with respect to the weights and the delays. One problem with the approach taken is that sometimes the network is unable to converge to a proper separating solution. It is therefore crucial for convergence that the initial delay estimates are approximately in the correct range, which can be achieved by using prior knowledge of the recording situation. Torkkola (1996a) also extends the algorithm to the echoic case.

The simultaneous diagonalisation technique of the SOBI algorithm has also been transposed to the anechoic case. Yeredor (2001) uses joint diagonalisation of spectral matrices to estimate the mixing coefficients as well as the delays. An analysis is provided in the frequency domain, resulting in a diagonalisation of a set of spectral matrices at different selected frequencies. The problem is to find a diagonalising matrix \mathbf{B} , which can be factored into a matrix of mixing coefficients and delays. This problem can be solved by extending Yeredor (2000), which solves the joint diagonalisation problem by alternately minimising a least squares cost function with respect to a matrix of mixing coefficients \mathbf{A} , delays \mathbf{D} , and source spectra $\mathbf{\Gamma}$. The algorithm is iterated over three steps: first minimise the least squares cost function with respect to $\mathbf{\Gamma}$ with \mathbf{A} and \mathbf{D} fixed; then minimise

with respect to each column of \mathbf{A} , assuming the other columns and \mathbf{T} and $\mathbf{\Gamma}$ are fixed and finally minimise with respect to each column of \mathbf{D} , assuming the other columns and \mathbf{A} and $\mathbf{\Gamma}$ are fixed. The parameter estimates after convergence define the mixing process.

C. Echoic Mixing. Lambert (1996) develops an FIR matrix algebra that can be used to solve problems involving echoic mixtures. The idea is to extend the algebra of scalar matrices to matrices of filters (time domain) or polynomials (frequency domain). Lambert presents techniques that compute the inverse of a matrix of FIR filters. These techniques are required to determine the inverse mixing process and have enjoyed wide deployment in separation methods for echoic mixtures.

Westner and Bove (1999) discuss the separation of over-determined mixtures (more sensors than sources). A technique is presented that involves taking the impulse response of the room in which the observations are made and then parameterising the FIR filters in the mixing process with the inverse of the impulse response. As described in Lambert (1996), it is possible to use standard scalar matrix operations to invert FIR polynomial matrices, and in the case of an over-determined system, it is possible to use a pseudo-inverse to determine the unmixing matrix. The motivation for using more sensors than sources is quality of separation: results are presented that show that the signal to noise ratio of the separated sources improves with an increase in the number of sensors. This approach, where impulse responses are measured, diverges from the blind separation model. However this technique indicates the information needed to estimate sources from echoic mixtures.

A computationally efficient method to determine the inverse mixing process is to estimate the parameters of the mixing process in the frequency domain, where convolution becomes scalar multiplication, $f(t) * g(t) \xrightarrow{\mathcal{F}} f(\omega) g(\omega)$. It is possible to transform an FIR filter matrix to an FIR polynomial matrix by performing a frequency transform on its elements. FIR polynomial matrices are matrices whose elements are complex valued polynomials (Lambert, 1996). A consequence of the transformation is that the frequency components observed at the sensors are in fact instantaneous mixtures of the original frequency components of the sources, i.e each frequency bin of $\mathbf{x}(t)$ is an instantaneous mixture. The mixing process can be observed in the frequency domain by performing an FFT transformation on $\mathbf{x}(t)$. When the mixing process is transformed in this way, the resultant FIR polynomial matrix can be modelled by a combination of multiple complex valued matrices and therefore can be viewed as solving multiple instantaneous mixing problems in the frequency domain.

Smaragdis (1998) uses this approach and presents an algorithm that employs BS-Infomax. The BS-Infomax learning rules are reformulated for complex domain data and are used to learn the unmixing process at each frequency bin. One major problem with this approach is that of permutation and scaling: the columns of the resultant mixing matrices may not all correspond to the same source, which results in incorrect source estimation. Although a number of heuristics to alleviate this problem are presented (Ane-müller and Kollmeier, 2003), it remains an open problem. A flow diagram of this algorithm is shown in Figure 7.

Murata et al. (2001) also formulated the problem of convolutive mixture separation in the frequency domain. The instantaneous mixtures at each frequency bin are separated using simultaneous diagonalisation similar to the SOBI algorithm. To solve the permu-

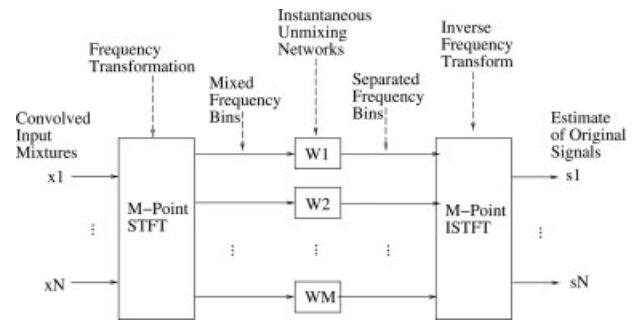


Figure 7. Flow diagram of the Smaragdis' frequency domain algorithm for echoic separation. (From Smaragdis, 1998, Figure 3, with permission).

tation problem, a method based on the temporal structure of signals that exploits the non-stationarity of speech is introduced. The separated components at each bin are combined with the components in the other bins that exhibit a similar envelope over some specified time. This method makes use of the fact that non-stationary signals exhibit correlation between the frequency bins of the signals spectrum.

Sahlin and Broman (1998) present an algorithm that is applied to the problem of separating a signal from noise. A mobile phone is fitted with two microphones and the noise performance of the algorithm is tested in a range of echoic environments. Echoic separation is formulated as an optimisation problem in which the cost function is the sum of squared cross correlations between the separated signals, the goal is to estimate appropriate filter weights that will achieve separation. Minimisation of the cost function is performed using a *Gauss-Newton* search, and when separation is achieved, the separated sources are mutually uncorrelated. One problem that arises in convolutive mixture separation is to determine the filter order required for separation. One way to tackle this problem is to use a technique known as 'leaking' (Ljung and Sjöberg, 1992), which involves modifying the cost function with the benefit that there is no performance degradation because of over-parameterisation of the algorithm.

Convolutive mixtures also appear in digital information transmission. A typical assumption in this area is that the number of sources is less than the number of observations. Icart and Gautier (1996) present a method in which the generative model of Eq. (5) is modelled using an *auto-regressive moving average* (ARMA) model. The first stage of the algorithm is to estimate the convolutive mixture using a linear predictor, which is used to create a filter that provides an instantaneous mixture of the source signals. Sources can then be estimated from this new mixture using any instantaneous mixture method. A similar approach is also presented in Gorokhov et al. (1996).

The problem of blind separation of convolutive sources can be compared with that of blind channel equalisation (sometimes called blind deconvolution). Blind equalisation is an adaptive filtering technique, which can restore a signal corrupted by a multipath channel. Lambert (1995) extended this idea to the multichannel problem of blind separation where separation is achieved by decorrelating multiple source signals. Lambert formulates a multichannel blind equalisation cost function and makes use of the Bussgang property (Bussgang, 1952) (also discussed in Fiori (2002)) to achieve separation (Lambert and Bell, 1997). The Bussgang prop-

erty can be stated as follows: two random variables x_i and x_j are independent if and only if $E[f(x_i)g(x_j)] = E[f(x_i)]E[g(x_j)]$ for all functions $f(\cdot)$ and $g(\cdot)$.

Baxter and McWhirter (2003) present an algorithm, which is a generalisation of instantaneous ICA, where hidden rotation matrix estimation is replaced by estimation of a Hidden Paraunitary Matrix (HPM). An initial process of strong decorrelation and spectral equalisation (whitening), based entirely on second order statistics, is followed by the identification of a HPM that necessitates the use of fourth order statistics. The concept of an Elementary Paraunitary Matrix (EPM) is introduced, and this takes the form of a number of delays applied to one of the channels, followed by a rotation. The HPM is built up as a sequence of EPMs. The parameters required for the HPM are determined using a novel method termed the sequential best rotation algorithm. At each iteration of the algorithm, a delay and rotation angle are computed and between them they specify an EPM. These parameters are chosen to maximise statistical independence (using any instantaneous ICA method) at the output of that stage. The number of stages is not specified in advance. Instead, the algorithm continues until no further improvement can be achieved, which differs from previous methods (Vaidyanathan, 1993) of paraunitary matrix estimation.

Monastral source separation of convolutive sources has also received some attention. A biologically inspired technique that exploits spectral cues is presented in Pearlmutter and Zador (2004). When sound reaches an organism's inner ear, the sound's spectrum is 'coloured' by the head and the shape of the ears. This spectral colouring or filtering is known as the *head-related transfer function* (HRTF) and is defined by the direction of the sound and the acoustic properties of the ear. When using the HRTF, it is assumed that each source has a unique position in space. The auditory scene is segregated into a number of different locations each with an associated HRTF filter. Sources coming from these locations will be coloured by its associated filter, which indicates the source's position; thus, identification of the HRTF filter applied to each source will lead to separation of sources. This can be achieved by sparsely representing the sources in an over-complete (under-determined) dictionary, which is performed as follows: a known HRTF is used with a given signal dictionary where each dictionary element is filtered by the HRTF filter at each location in the auditory scene, resulting in an over-complete signal dictionary. The monaural mixture signal is then decomposed into these dictionary elements by *L_1 -norm minimisation* (see Section IVA) and the resultant coefficients indicate the location of the sources in the signal. The dictionary elements of the locations identified are then scaled by their coefficients and linearly combined to create the estimated sources. By estimating the coefficients by using a post-HRTF (sensor space) dictionary but reconstructing using a pre-HRTF (source space) dictionary, separation and deconvolution can be simultaneously achieved.

Careful placement of microphones can greatly simplify the problem of separating echoic sources. Sanchis et al. (2004) discuss how to convert echoic mixtures to instantaneous mixtures using two stereophonic microphones: the microphones are positioned at the same point in space (making the observations instantaneous at each microphone) but in different directions. The directional properties of the microphones cause the mixtures observed at the microphones to have a differing contribution from each source (if regular non-directional microphones were used, each microphone would record the same observation). Thus, the problem is converted to the instan-

taneous case. In some practical situations, the mixtures are not truly instantaneous because of reverberations and are considered to be still echoic, and the advantage of the microphone arrangement in this case is that fewer taps are need for the separation filters. Similar work is presented by Katayama et al. (2004).

IV. SEPARATION TECHNIQUES

Subsequent to the estimation of the mixing matrix, separation of the underlying sources can be performed. The complexity of the separation process is influenced by the mixing model used and the relative number of sources and sensors. The next section presents a number of techniques used in the separation stage of BSS algorithms.

A. Instantaneous Separation. Separation in the even-determined case can be achieved by a linear transformation using the unmixing matrix \mathbf{W} ,

$$\hat{\mathbf{s}}(t) = \mathbf{W}\mathbf{x}(t), \quad (8)$$

where $\hat{\mathbf{s}}(t)$ holds the estimated sources at time t , $\hat{\mathbf{A}}$ is the estimated mixing matrix and $\mathbf{W} = \hat{\mathbf{A}}^{-1}$ up to permutation and scaling of the rows.

For the under-determined case, which is usually restricted to sparse methods (see Table I), a linear transformation is not possible, since $\hat{\mathbf{A}}\hat{\mathbf{s}}(t) = \mathbf{x}(t)$ has more unknowns in \mathbf{s} than knowns in \mathbf{x} and is therefore non-invertible. Consequently, some non-linear technique is needed to estimate the sources. Usually, these techniques involve assigning observed data $\mathbf{x}(t)$ to the columns of $\hat{\mathbf{A}}$ that characterise each source. The most rudimentary technique is to hard assign each data point to only one source based on some measure of proximity to columns of $\hat{\mathbf{A}}$ (Vielva et al., 2000, 2002; Lin et al., 1997). A logical extension of the previous technique is the partial assignment of each data point to multiple sources. This is generally done by minimisation of the L_1 -norm (sometimes referred to as the shortest-path algorithm (Bofill and Zibulevsky, 2000) or basis pursuit (Chen et al., 1998)). L_1 -norm minimisation is a piecewise linear operation that partially assigns the energy of $\mathbf{x}(t)$ to the M columns of $\hat{\mathbf{A}}$ that form a cone around $\mathbf{x}(t)$ in \mathbb{R}^M space. The remaining $N - M$ columns are assigned zero coefficients; therefore, L_1 -norm minimisation provides more accurate source estimates when the sources are sparse, in this case, when the number of sources active at any one time is less than or equal to M . L_1 -norm minimisation can be accomplished by formulating the problem as a linear program

$$\text{minimise } \|\hat{\mathbf{s}}(t)\|_1 \text{ subject to } \hat{\mathbf{A}}\hat{\mathbf{s}}(t) = \mathbf{x}(t), \quad (9)$$

where the observed data point $\mathbf{x}(t)$ is in an appropriate domain and the $\hat{\mathbf{s}}(t)$ coefficients, properly arranged, constitute the estimated sources, $\hat{\mathbf{s}} = [\hat{s}(1) \dots \hat{s}(T)]$. A detailed discussion of signal recovery using L_1 -norm minimisation is presented by Takigawa et al. (2004). When using complex data, as in the case of a FFT representation, the real and imaginary parts are treated separately; thus, doubling the number of coefficients.

Source separation can be formalised in a probabilistic framework: given $\mathbf{x}(t)$ and $\hat{\mathbf{A}}$, a standard approach for reconstructing $\hat{\mathbf{s}}(t)$ is maximum likelihood, which finds an $\hat{\mathbf{s}}(t)$ that maximises the posterior probability. If the prior probability $P(\hat{\mathbf{s}}(t))$ is Laplacian

($P(\hat{\mathbf{s}}(t)) \propto e^{-|\hat{\mathbf{s}}(t)|}$), then this probabilistic approach becomes the L_1 -norm minimisation of Eq. (9)

$$\begin{aligned}
\hat{\mathbf{s}}(t) &= \max_{\hat{\mathbf{A}}\hat{\mathbf{s}}(t)=\mathbf{x}(t)} P(\hat{\mathbf{s}}(t)|\mathbf{x}(t), \hat{\mathbf{A}}) \\
&= \max_{\hat{\mathbf{A}}\hat{\mathbf{s}}(t)=\mathbf{x}(t)} P(\mathbf{x}(t)|\hat{\mathbf{s}}(t), \hat{\mathbf{A}})P(\hat{\mathbf{s}}(t)) \\
&= \max_{\hat{\mathbf{A}}\hat{\mathbf{s}}(t)=\mathbf{x}(t)} P(\hat{\mathbf{s}}(t)) \\
&= \max_{\hat{\mathbf{A}}\hat{\mathbf{s}}(t)=\mathbf{x}(t)} \exp(-(|\hat{\mathbf{s}}_1(t)| + \dots + |\hat{\mathbf{s}}_N(t)|)) \\
&= \min_{\hat{\mathbf{A}}\hat{\mathbf{s}}(t)=\mathbf{x}(t)} |\hat{\mathbf{s}}_1(t)| + \dots + |\hat{\mathbf{s}}_N(t)| \\
&= \min_{\hat{\mathbf{A}}\hat{\mathbf{s}}(t)=\mathbf{x}(t)} \|\hat{\mathbf{s}}(t)\|_1.
\end{aligned}$$

B. Anechoic Separation. For algorithms that are constrained to the even-determined case (Yeredor, 2001), separation can be achieved by linear transform in the frequency domain. Observations are transformed using a DFT and the underlying sources are estimated by multiplying with the inverse product of the amplitude and delay matrix, and the sources are then transformed back to the time domain.

For algorithms such as DUET (Jourjine et al., 2000) that use time–frequency representations of the observations, one powerful approach, known as *time–frequency masking* (Yilmaz and Rickard, 2004), is to partition the time–frequency domain into regions corresponding to individual sources. These regions are defined by proximity to the estimated delay and amplitude parameters and each source is then estimated by synthesising the coefficients contained in its region. This technique relies on the assumption of source disjointness; sources are said to be disjoint if the intersection of the supports of the sources is zero, i.e. each source occupies an exclusive region in the time–frequency domain. Although this condition is not fully met in practice, particularly in acoustics, it has proven a very useful and practical approximation (Rickard and Yilmaz, 2002).

Bofill (2002) presents a method that reduces the anechoic case to the instantaneous case, resulting in individual matrices for the amplitudes and delays. Sources can then be estimated using methods presented in the previous subsection. Bofill uses a technique similar to L_1 -norm minimisation called *second order conic programming*, the difference being that the latter can include the magnitudes of complex numbers in the L_1 -norm; therefore, eliminating the need for, and approximation made by, putting the real and imaginary components separately into the L_1 -norm.

C. Echoic Separation. Separation of echoic mixtures is generally only attempted in the even-determined case, where it can be achieved by convolving mixture observations with an inverse mixing process \mathbf{W} ,

$$\hat{\mathbf{s}}(t) = \mathbf{W} * \mathbf{x}(t). \quad (10)$$

The most widely deployed method to obtain an inverse mixing process is the FIR matrix inversion of Lambert (1995). Matrices of filters have inverses of the same form as matrices of reals, and a

convolutive mixing process with two mixtures and two sources would have an inverse of the form

$$\mathbf{W} = \hat{\mathbf{A}}^{-1} = \frac{1}{\tilde{A}_{11} * \tilde{A}_{22} - \tilde{A}_{12} * \tilde{A}_{21}} \begin{bmatrix} \tilde{A}_{22} & -\tilde{A}_{12} \\ -\tilde{A}_{21} & \tilde{A}_{11} \end{bmatrix}. \quad (11)$$

Thus, in the low-noise even-determined case, estimation of the convolutive mixing process allows the construction of an unmixing process.

For algorithms such as Icart and Gautier (1996) and Gorokhov et al. (1996), which reduce the problem to one of instantaneous mixing, the methods described in Sections IIIA and IVA are sufficient to unmix the original convolved sources. For the blind equalisation approach of Lambert (1995), source estimates can be obtained by making use of feedback where the current best estimates of \tilde{A}_{21} and \tilde{A}_{12} are used to obtain estimates of the sources. The equations for the two sensor two source case are as follows:

$$\begin{aligned}
\hat{s}_1(t) &= x_1(t) - \hat{s}_2(t) * \tilde{A}_{21}, \\
\hat{s}_2(t) &= x_2(t) - \hat{s}_1(t) * \tilde{A}_{12}.
\end{aligned} \quad (12)$$

V. JOINT ESTIMATION

The previous two sections outlined a staged approach to BSS, where source separation is performed after estimating the mixing process. It is also possible to estimate the sources and mixing process concurrently by selecting an appropriate cost function and defining the problem as an optimisation problem. Such a cost function can be constructed using a measure of distance between \mathbf{X} and the product $\mathbf{A}\mathbf{S}$. One typical measure is the square of the Euclidean distance between \mathbf{X} and $\mathbf{A}\mathbf{S}$.

$$C(\mathbf{X}, \mathbf{A}, \mathbf{S}) = \|\mathbf{X} - \mathbf{A}\mathbf{S}\|^2 = \sum_{ij} (x_{ij} - (\mathbf{A}\mathbf{S})_{ij})^2. \quad (13)$$

This is insufficient to fully constrain the solution, since an optimum can be transformed into another equivalent optimum by $\mathbf{S} \mapsto \mathbf{U}\mathbf{S}$, $\mathbf{A} \mapsto \mathbf{A}\mathbf{U}^{-1}$, where \mathbf{U} is any invertible matrix. However, we are now in a position to formulate $\mathbf{X} = \mathbf{A}\mathbf{S}$ as an optimisation problem. We will focus our discussion on the emerging field of *non-negative matrix factorisation* (NMF) (Lee and Seung, 1999), which is an optimisation problem of the following form

$$\text{minimize } C(\mathbf{X}, \mathbf{A}, \mathbf{S}) \text{ subject to } \mathbf{A}, \mathbf{S} \geq 0.$$

The presented cost function is convex in \mathbf{A} or \mathbf{S} only, but not in both variables together. Therefore, it is unrealistic to expect an algorithm to find a global minima for both variables together and usually, algorithms alternate updates of \mathbf{A} and \mathbf{S} .

A. Non-Negative Matrix Factorisation. Non-negative matrix factorisation is a technique for the decomposition of multivariate data. It is based on the fact that in many data processing tasks, negative numbers are physically meaningless and contradict physical realities. For example, when considering grey-scale images, the principal component analysis decomposition of an image may result in basis vectors that have both positive and negative components. This image is then represented by a linear combination of these basis vectors weighted by both positive and negative coefficients, with some basis vectors being cancelled out by others. Since negative basis has no real-world representation in a grey-scale image context, this has led researchers to argue that the search for a basis should be

confined to a non-negative basis with real-world manifestations. Formally, this idea can be interpreted as decomposing a non-negative matrix \mathbf{P} into two non-negative factors \mathbf{V} and \mathbf{H} such that

$$\mathbf{P} \approx \mathbf{V}\mathbf{H}.$$

In a source separation context, provided that there is a non-negative constraint on \mathbf{X} , \mathbf{S} and \mathbf{A} , the above can be replaced by

$$\mathbf{X} \approx \mathbf{A}\mathbf{S}. \quad (14)$$

Although an exact decomposition (i.e. $\mathbf{X} = \mathbf{A}\mathbf{S}$) is possible, in the context of source separation reconstructions are usually approximate. The significance of Eq. (14) is that when rewritten column by column, i.e. $\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$, it is evident that $\mathbf{x}(t)$ is approximated by a linear combination of the columns of \mathbf{A} weighted by the components of $\mathbf{s}(t)$. Therefore, \mathbf{A} can be regarded as having an optimised basis for the approximation of the data in \mathbf{X} . Since \mathbf{A} contains N basis vectors in its columns to represent T data vectors in \mathbf{X} , where typically $T \gg N$, good approximation can only be achieved if the basis vectors discover structure that is latent in the data. For a sparse non-negative representation of \mathbf{X} , provided that the dimensions of \mathbf{A} are appropriate, the latent structure of the data may be defined by a mixing process and in that case \mathbf{S} will contain estimates of the original sources in its rows. It is worth noting that a non-negativity constraint on the mixing process defined by \mathbf{A} limits the applicability of NMF in source separation, as the mixing matrix can often include negative components. However, NMF readily lends itself to monaural source separation problems.

For data such as speech, which contain negative components, the power spectrum density can be calculated by DFT of the data, $\mathbf{X}(t) \mapsto |\mathbf{X}(w)|^2$, and this non-negative representation can then be factorised.

A.1. NMF Algorithms. In a highly influential paper, Lee and Seung (2001) presented an algorithm that optimises \mathbf{A} and \mathbf{S} individually based on the cost function of Eq. (13). This is the standard algorithm for NMF and works as follows: when optimising \mathbf{S} , \mathbf{A} is fixed and a multiplicative update rule is used to update \mathbf{S} , and vice versa,

$$\mathbf{S}^{t+1} = \mathbf{S}^t \frac{\mathbf{A}^T \mathbf{X}}{\mathbf{A}^T \mathbf{A} \mathbf{S}^t}, \quad (15)$$

$$\mathbf{A}^{t+1} = \mathbf{A}^t \frac{\mathbf{X} \mathbf{S}^t}{\mathbf{A} \mathbf{S}^t \mathbf{S}^t}. \quad (16)$$

It is important to have an understanding of the conditions required for a correct NMF solution. Donoho and Stodden (2004) investigate what conditions make the NMF objective well defined and the answer correct independent of the algorithm used. Non-negative matrix factorisation can be interpreted geometrically as finding its geometric counterpart called the simplicial cone. The developed geometric conditions are considered in the context of image articulation libraries and it is shown that in the case of separable factorial articulation families, non-negative matrix factorisation is essentially unique. An experiment using the Lee and Seung (2001) algorithm is presented, and the results demonstrate that the theoretical results are predictive of the performance of the algorithm.

As discussed in Section I, it is often desirable to decompose data into a sparse representation. In the context of NMF, the requirement for a sparse decomposition of \mathbf{X} would result in another constraint being added to the optimisation problem. Non-negative matrix factorisation with an added sparseness constraint is presented in Hoyer (2002), where the sparse decomposition technique introduced is referred to as *non-negative sparse coding* (NSC). This technique uses the same cost function as Eq. (13) but introduces an additional term that enforces sparseness on \mathbf{S} ,

$$C(\mathbf{X}, \mathbf{A}, \mathbf{S}) = \frac{1}{2} \|\mathbf{X} - \mathbf{A}\mathbf{S}\|^2 + \lambda \sum_{ij} f(s_{ij}), \quad (17)$$

where the function $f(\cdot)$ defines how sparseness is measured (typically $f(\xi) = |\xi|$) and the parameter λ controls the trade off between sparseness and accurate reconstruction. This objective creates a new problem: $f(\cdot)$ is typically a strictly increasing function of the absolute value of its argument, so it is possible that the objective can be decreased by scaling up \mathbf{A} and scaling down \mathbf{S} ($\mathbf{A} \mapsto \alpha\mathbf{A}$ and $\mathbf{S} \mapsto (1/\alpha)\mathbf{S}$, with $\alpha > 1$). This situation does not alter the first term in the objective function, but will cause the second term to decrease, resulting in the elements of \mathbf{A} growing without bound and \mathbf{S} tending toward zero. The consequence is that the solution arrived at by the optimisation algorithm is not influenced by the second term of the objective function and the resultant \mathbf{S} matrix is not sparse. Therefore, another constraint needs to be introduced in order to make the cost function well-defined. This is achieved by fixing the norm of the columns of \mathbf{A} to unity, which constrains the scale of the elements in \mathbf{A} and \mathbf{S} .

The algorithm presented by Hoyer (2002) is similar to Lee and Seung (2001), although the unit-norm constraint on the columns of \mathbf{A} complicates the update calculation for \mathbf{A} ; the update for \mathbf{A} is

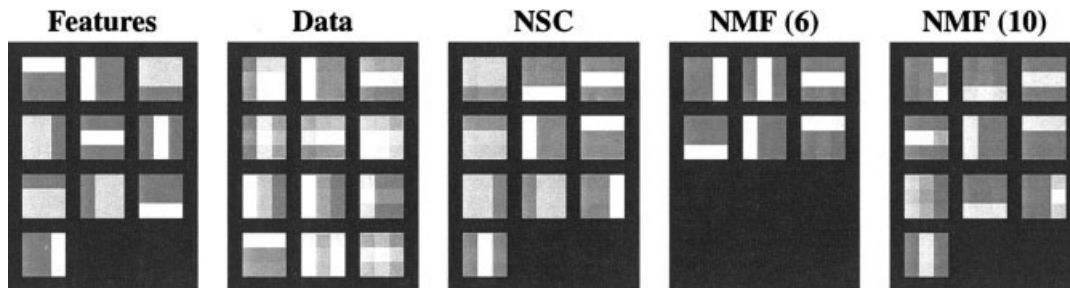


Figure 8. Experiments on bars data. Features: the 10 original features (columns of \mathbf{A}_{orig}) that were used to construct the dataset. Data: a random sample of 12 data vectors (columns of \mathbf{X}). These constitute superpositions of the original features. NSC: features (columns of \mathbf{A}) learned by NSC, with dimensionality of the hidden representation (number of rows of \mathbf{S}) equal to 10, starting from random initial values. NMF (6): features learned by NMF, with dimensionality 6. NMF (10): features learned by NMF, with dimensionality 10. (From Hoyer, 2002, Figure 1, with permission).

replaced by a projected gradient descent calculation with the resultant update elements set to zero if negative and each column normalised. An experiment using simple 3×3 pixel images in which each image is treated as a 9-element vector is presented (see Fig. 8). A data matrix \mathbf{X} with 12 data images in its columns is constructed from random non-negative data \mathbf{S}_{orig} using the generative model $\mathbf{X} = \mathbf{A}_{\text{orig}} \mathbf{S}_{\text{orig}}$. An over-complete basis of 10 images is manually constructed from a complete basis of 6 images with the remaining 4 images constructed from a combination of 2 from the complete basis, these images are then placed in the columns of \mathbf{A}_{orig} . NSC is compared with NMF and the results show that, provided the dimensions of \mathbf{A} are appropriate, NSC correctly identifies all the over-complete basis vectors of the data set, while NMF identifies only the complete basis vectors. The images in \mathbf{X} can be perfectly described by the complete subset of the over-complete basis vectors, and the advantage of identifying the entire over-complete basis is that a sparser description of the data can be achieved.

An interesting application of NSC for monaural sound separation using a temporal continuity assumption is presented in Virtanen (2003). The Temporal continuity assumption is motivated by the human auditory system (Bregman, 1990) and assumes that the mixture sources have constant spectra with time-varying gain. This algorithm considers only monaural signals; therefore, a different generative model is used. The observed signal is split into T frames of F samples, the power spectrum within each frame is calculated using a DFT and these frames are adjoined to form the columns of \mathbf{X} . The matrix of constant source spectra \mathbf{S} has $N \times F$ dimensions with spectra contained in each row. \mathbf{A} is a $T \times N$ matrix containing time-varying spectral gain information and defines the mixing of the constant spectra in \mathbf{S} for each frame. It is also assumed that the sound sources are inactive most of the time resulting in a sparse \mathbf{A} matrix. The observations \mathbf{X} are generated by $\mathbf{X} = \mathbf{A}\mathbf{S}$. The additional temporal continuity requirement can be included as an additional constraint on the NSC cost function of Eq. (17)

$$C(\mathbf{X}, \mathbf{A}, \mathbf{S}) = w_1 \|\mathbf{X} - \mathbf{A}\mathbf{S}\|_2^2 + w_2 \|\mathbf{A}\|_1 + w_3 \sum_{t=1}^T \sum_{i=1}^M |a_{t-1,i} - a_{t,i}|, \quad (18)$$

where w_1 , w_2 and w_3 are constants of proportionality that balance the cost function for accuracy of reconstruction, sparseness and temporal continuity, respectively. Temporal continuity is achieved by the third term of Eq. (18), which ensures that the gain differences between consecutive frames are minimised. It is also worth noting that the sparse constraint has shifted from \mathbf{S} to \mathbf{A} . The optimisation algorithm used is similar to Hoyer (2002) and is applied to the task of separation and transcription of drum sources from real-world polyphonic music signals. A similar method that makes use of temporal structure in a different way to that described above is presented by Smaragdis (2004).

Non-negative constraints have also found their way into a traditional ICA framework. Plumbley (2003) presents an algorithm that places a constraint of non-negativity on the sources. It is assumed that observations are noiseless and sources are *well-grounded*, i.e. they have a non-vanishing probability density function in the neighbourhood of zero. It is shown that the combination of pre-whitening and calculation of an appropriate orthonormal rotation is both necessary and sufficient for finding the set of underlying sources, provided that the sources are well-grounded. Two cost functions are

presented that can be used to determine the rotation matrix; these cost functions can then be used by any suitable optimisation algorithm. One interesting insight that emerges is that there is no requirement for the unwhitened inputs and mixing matrix to be non-negative in contrast with NMF. Although the discussion concentrates on sources with well-grounded distributions, it is expected that the algorithm will work particularly well with sparse distributions. This method diverges from the joint estimation approach because sources are estimated by a linear transformation, but illustrates how non-negative matrix constraints can be introduced into an ICA framework. Non-negativity constraints have also been applied to ICA for hyperspectral image analysis (Parra et al., 2000).

VI. FURTHER READING

The authors would like to recommend the following books that they have found useful in their research. For a detailed discussion on ICA techniques and principles, consult Roberts and Everson (2001) and Hyvärinen et al. (2001). For an introduction to blind deconvolution and associated algorithms including Bussgang techniques, consult Haykin (1994). A modern overview of the different approaches taken (including psychology, physiology, engineering and computer science) for the cocktail party problem is presented in Divenyi (2005). For background material on blind signal processing including multipath blind deconvolution, refer to Cichocki and Amari (2002). For a solid entry-level introduction to machine learning principles and techniques, consult Duda et al. (2001).

VII. CONCLUSION

The problem of BSS is a challenging and interesting one, which has enjoyed much attention over the last two decades. The complexity of the problem is influenced by environmental considerations, the number of sensor observations, and the number of underlying sources to be separated. This survey presented an overview of the techniques used and the different approaches taken to solve the problem. The problem can be solved by a staged approach in which the mixing process is estimated first and the sources are estimated subsequently. Alternatively, the problem can be viewed as an optimisation problem in which the mixing process and source estimates are estimated jointly. The adoption of an assumption of sparsity, by way of a suitable basis transformation, ameliorates the problem somewhat by enabling algorithms to efficiently exploit the structure of the mixing process. The sparse assumption also provides for a solution to the under-determined case. It is envisaged that the ability of sparse representations to efficiently represent useful statistical structure in a signal will lead to an increased number of practical signal processing algorithms in which sparse representations play a key role.

ACKNOWLEDGMENTS

Higher Education Authority of Ireland (An tÚdarás Um Ard-Oideachas) and Science Foundation Ireland grant 00/PI.1/C067.

REFERENCES

- S. Amari, A. Cichocki, and H.H. Yang, A new learning algorithm for blind signal separation, In *Advances in neural information processing systems*, Vol. 8, MIT Press, Cambridge, MA, 1996.
- J. Anemüller and B. Kollmeier, Adaptive separation of acoustic sources for anechoic conditions: A constrained frequency domain approach, *Speech Commun* 39 (2003), 79–95.

- M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones, *Acoust Sci Technol* 22 (2001), 149–157.
- A.D. Back and A.S. Weigend, A first application of independent component analysis to extracting structure from stock returns, *Int J Neural Syst* 8 (1997), 473–484.
- D. Barry, B. Lawlor, and E. Coyle, Sound source separation: Azimuth discrimination and resynthesis, In *Proc 7th Int Conf on Digital Audio Effects (DAFX-04)*, 2004, pp. 240–244.
- P.D. Baxter and J.G. McWhirter, Blind signal separation of convolutive mixtures, In *The Thirty-Seventh Asilomar Conf on Signals, Systems and Computers*, 2003, pp. 124–128.
- A.J. Bell and T.J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, *Neural Comput* 7 (1995), 1129–1159.
- A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and É. Moulines, A blind source separation technique based on second order statistics, *IEEE Trans Signal Proc* 45 (1997), 434–444.
- A. Belouchrani and J.-F. Cardoso, Maximum likelihood source separation for discrete sources, In *Proc EUSIPCO*, 1994, pp. 768–771.
- P. Bofill, Underdetermined blind separation of delayed sound sources in the frequency domain. Technical Report UPC-DAC-2001-14, Universitat Politècnica de Catalunya, Barcelona, 2002.
- P. Bofill and M. Zibulevsky, Blind separation of more sources than mixtures using the sparsity of the short-time Fourier transform, In *2nd Int Workshop on Independent Component Analysis and Blind Signal Separation*, Helsinki, Finland, June 19–20, 2000, pp. 87–92.
- P. Bofill and M. Zibulevsky, Underdetermined blind source separation using sparse representations, *Signal Process* 81 (2001), 2353–2362.
- A.S. Bregman, *Auditory scene analysis: The perceptual organization of sound*, MIT Press, Cambridge, MA, 1990. ISBN 0-262-02297-4.
- J.J. Busgang, Crosscorrelation functions of amplitude-distorted Gaussian signals, Technical Report 216, MIT Research Laboratory of Electronics, Cambridge, MA, 1952.
- J.-F. Cardoso, Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem, In *Proc IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP'90)*, Albuquerque, NM, 1990, pp. 2655–2658.
- J.-F. Cardoso, Infomax and maximum likelihood for blind source separation *IEEE Signal Process Lett* 4 (1997), 112–114.
- J.-F. Cardoso, Blind signal separation: Statistical principles, *Proc of the IEEE*, 9(10) (1998), 2009–2025. Available at <ftp://sig.enst.fr/pub/jfc/Papers/Proc IEEE.us.ps.gz>.
- J.-F. Cardoso, J. Delabrouille, and G. Patanchon, Independent component analysis of the cosmic microwave background, In *Fourth Int Symp on Independent Component Analysis and Blind Signal Separation*, Nara, Japan, April 1–4, 2003, pp. 1111–1116.
- J.-F. Cardoso and A. Souloumiac, Blind beamforming for non-Gaussian signals, *IEE Proc F* 140(6) (1993), 362–370. Available at <ftp://tsi.enst.fr/pub/jfc/Papers/iee.ps.gz>.
- S.S. Chen, D.L. Donoho, and M.A. Saunders, Atomic decomposition by basis pursuit, *SIAM J Sci Comput* 20 (1998), 33–61.
- C.E. Cherry, Some experiments in the recognition of speech, with one and two ears, *J Acoust Soc Am* 25 (1953), 975–979.
- C. Choi, Real-time binaural blind source separation, In *Fourth Int Symp on Independent Component Analysis and Blind Signal Separation*, Nara, Japan, April 1–4, 2003, pp. 567–572.
- A. Cichocki and S.-I. Amari, *Adaptive blind signal and image processing: Learning algorithms and applications*, John Wiley & Sons, New York, 2002. ISBN 0-471-60791-6.
- P. Comon, Independent component analysis: A new concept, *Signal Process* 36 (1994), 287–314.
- G. Darrois, Analyse générale des liaisons stochastiques, *Rev Inst Int Stat* 21 (1953), 2–8.
- A.P. Dempster, N.M. Laird, and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J R Stat Soc Ser B* 39 (1976), 1–38.
- M.R. DeWeese, M. Wehr, and A.M. Zador, Binary spiking in auditory cortex, *J Neurosci* 23 (2003), 7940–7949.
- P. Divenyi (Editor). *Speech separation by humans and machines*, Kluwer Academic Publishers, Boston, 2005. ISBN 1-4020-8001-8.
- D. Donoho and V. Stodden, When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in neural information processing systems* 16. MIT Press, Cambridge, MA, 2004. Available at http://books.nips.cc/papers/files/nips16/NIPS2003_LT10.pdf.
- J. Du, C.-H. Lee, H.-K. Lee, and Y.-H. Suh, BSS: A new approach for watermark attack, In *Proc 4th Int Symp on Multimedia Software Engineering (ISMSE 2002)*, 2002, pp. 182–187.
- R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern classification*, 2nd edition, John Wiley & Sons, New York, 2001.
- S. Fiori, Notes on cost functions and estimators for ‘Busgang’ adaptive blind equalization, *Eur Trans Telecommun (ETT)* 13 (2002), 631–634.
- P. Földiák and M. Young, Sparse coding in the primate cortex, In *The handbook of brain theory and neural networks*, MIT Press, Cambridge, MA, 1995, pp. 895–898.
- M. Gaeta and J.-L. Lacoume, Source separation without prior knowledge: The maximum likelihood solution, In *Proc EUSIPCO'90*, 1990, pp. 621–624.
- A. Gorokhov, P. Loubaton, and E. Moulines, Second order blind equalization in multiple input multiple output FIR systems: A weighted least squares approach, In *Proc IEEE Int Conf on Acoustics, Speech and Signal Processing*, Atlanta, GA, May 7–10 1996, pp. 2415–2418.
- S.S. Haykin, *Blind deconvolution*, In *Information and system sciences*, Prentice-Hall, Englewood Cliffs, NJ, 1994. ISBN 0-13-087362-4.
- J. Hérault and C. Jutten, Space or time adaptive signal processing by neural models, In *Proc AIP Conf on Neural Networks for Computing*, American Institute of Physics, 1986, pp. 206–211.
- P.O. Hoyer, Non-negative sparse coding, In *IEEE Workshop on Neural Networks for Signal Processing*, 2002.
- M.V. Hulle, Kernel-based equiprobabilistic topographic map formation, *Neural Comput* 10 (1998), 1847–1871.
- A. Hyvärinen, Survey on independent component analysis, *Neural Comput Surv* 2 (1999), 94–128.
- A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, John Wiley & Sons, New York, 2001.
- A. Hyvärinen and E. Oja, A fast fixed-point algorithm for independent component analysis, *Neural Comput* 9 (1997), 1483–1492.
- S. Icart and R. Gautier, Blind separation of convolutive mixtures using second and fourth order moments, In *Proc IEEE Int Conf on Acoustics, Speech, and Signal Processing (ICASSP'96)*, Atlanta, GA, USA, May 1996, pp. 3018–3021.
- A. Jourjine, S. Rickard, and O. Yilmaz, Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures, In *Proc IEEE Conf on Acoustics, Speech, and Signal Processing (ICASSP2000)*, June 2000, Vol. 5, pp. 2985–2988.
- T.-P. Jung, C. Humphries, T.-W. Lee, M.J. McKeown, V. Iragui, S. Makeig, and T.J. Sejnowski, Removing electroencephalographic artifacts by blind source separation, *Psychophysiology* 37 (2000), 163–178.
- T.-P. Jung, S. Makeig, M. Westerfield, J. Townsend, E. Courchesne, and T.J. Sejnowski, Analyzing and visualizing single-trial event-related potentials, In *Advances in neural information processing systems*, Vol. 11, MIT Press, Cambridge, MA, 1999, pp. 118–124.

- C. Jutten and J. Herault, Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture, *Signal Process* 24 (1991), 1–10.
- Y. Katayama, M. Ito, A.K. Barros, Y. Takeuchi, T. Matsumoto, H. Kudo, N. Ohnishi, and T. Mukai, Closely arranged directional microphone for source separation: Effectiveness in reduction of the number of taps and preventing factors, In *Fifth Int Conf on Independent Component Analysis*, Granada, Spain, Sept. 22–24, 2004, Lecture Notes in Computer Science, Vol. 3195, Springer-Verlag.
- M. Kearns, Y. Mansour, and A.Y. Ng, An information-theoretic analysis of hard and soft assignment methods for clustering. *Proc Thirteenth Conf on Uncertainty in Artificial Intelligence*, 1997, pp. 282–293.
- K.P. Körding, P. König, and D.J. Klein, Learning of sparse auditory receptive fields, In *Int Joint Conf on Neural Networks*, 2002. Available at <http://www.koerding.com/pubs/KKKijcnn01.pdf>.
- R.H. Lambert, A new method for source separation, In *Proc IEEE Conference on Acoustics, Speech, and Signal Processing*, Detroit, MI, 1995, pp. 2116–2119.
- R.H. Lambert, Multichannel blind deconvolution: fir matrix algebra and separation of multipath mixtures. Ph.D. thesis, Univ of Southern California, 1996.
- R. Lambert and A. Bell, Blind separation of multiple speakers in a multipath environment, In *Proc IEEE Int Conf on Acoustics, Speech, and Signal Processing (ICASSP'97)*, Munich, Germany, April 1997, pp. 423–426.
- D.D. Lee and H.S. Seung, Learning the parts of objects with nonnegative matrix factorization, *Nature* 401 (1999), 788–791.
- D.D. Lee and H.S. Seung, Algorithms for non-negative matrix factorization, In *Advances in neural information processing systems*, Vol. 13, MIT Press, Cambridge, MA, 2001 pp. 556–562. Available at citeseer.ist.psu.edu/lee00algorithms.html.
- T.-W. Lee, M.S. Lewicki, M. Girolami, and T.J. Sejnowski, Blind source separation of more sources than mixtures using overcomplete representations, *IEEE Signal Process Lett* 4 (1999), 87–90.
- M. Lewicki and T. Sejnowski, Learning overcomplete representations, In *Advances in neural information processing systems*, Vol. 10, MIT Press, Cambridge, MA, 1998, pp. 556–562.
- J.K. Lin, D.G. Grier, and J.D. Cowan, Feature extraction approach to blind source separation, In *IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, 1997, pp. 398–405.
- R. Linsker, An application of the principle of maximum information preservation to linear systems, In *Advances in neural information processing systems*, Morgan Kaufmann, San Francisco, CA, 1989, pp. 186–194.
- L. Ljung and J. Sjöberg, A comment on leaking in adaptive algorithms, In *4th IFAC Int Symp on Adaptive Systems in Control and Signal Processing, ACASP92*, 1992, pp. 377–382.
- J. MacQueen, Some methods for classification and analysis of multivariate observations, In L. M. L. Cam and J. Neyman (Editors), *Proc of the Fifth Berkeley Symp on Mathematical Statistics and Probability*, Berkeley, Vol. 1, University of California Press, 1967, pp. 281–297.
- M.J. McKeown, S. Makeig, G.G. Brown, T.P. Jung, S.S. Kindermann, A.J. Bell, and T.J. Sejnowski, Analysis of fMRI data by blind separation into independent spatial components, *Hum Brain Mapp* 6 (1998), 160–188.
- N. Murata, S. Ikeda, and A. Ziehe, An approach to blind source separation based on temporal structure of speech signals, *Neurocomputing* 41 (2001), 1–24.
- K. Nakadai, H.G. Okuno, and H. Kitano, Real-time sound source localization and separation for robot audition. In *Proc of 2002 Int Conf on Spoken Language Processing (ICSLP-2002)*, 2002, pp. 193–196.
- P.D. O’Grady and B.A. Pearlmutter, Hard-LOST: Modified k -means for oriented lines. In *Proc Irish Signals and Systems Conf*, Belfast, June 30–July 2, 2004a, pp. 247–252.
- P.D. O’Grady and B.A. Pearlmutter, Soft-LOST: EM on a mixture of oriented lines. In *Fifth Int Conf on Independent Component Analysis*, Granada, Spain, Sept. 22–24, 2004b, Lecture Notes In Computer Science, Vol. 3195, Springer-Verlag, pp. 430–436.
- L. Parra, K.-R. Mueller, C. Spence, A. Ziehe, and P. Sajda, Unmixing hyperspectral data, In *Advances in neural information processing systems*, Vol. 13, MIT Press, Cambridge, MA, 2000, pp. 942–948.
- L. Parra and P. Sajda, Blind source separation via generalized eigenvalue decomposition, *J Machine Learn Res* 4 (2003), 1261–1269.
- B.A. Pearlmutter and S. Jaramillo, Progress in blind separation of magnetoencephalographic data, In *Independent component analyses, wavelets, and neural networks*, Proceeding of the SPIE, Vol. 5102, 2003, pp. 129–134.
- B.A. Pearlmutter and L.C. Parra, A context-sensitive generalization of ICA, In *Int Conf on Neural Information Processing*, Hong Kong, Sept. 24–27, 1996, Springer-Verlag, pp. 151–157. Available at <http://www-bcl.cs.may.ie/~barak/papers/iconip-96-cica.ps.gz>.
- B.A. Pearlmutter and A.M. Zador, Monaural source separation using spectral cues, In *Fifth Int Conf on Independent Component Analysis*, Granada, Spain, Sept. 22–24, 2004, Lecture Notes in Computer Science, Vol. 3195, Springer-Verlag, pp. 478–485.
- K. Pearson, On lines and planes of closest fit to systems of points in space, *Philos Mag* 2 (1901), 559–572.
- D.T. Pham, P. Garrat, and C. Jutten, Separation of a mixture of independent sources through a maximum likelihood approach, In *European Signal Processing Conf*, 1992, pp. 771–774.
- J.C. Platt and F. Faggin, Networks for the separation of sources that are superimposed and delayed, In *Advances in neural information processing systems*, Vol. 4, Morgan Kaufmann, San Francisco, CA, 1992, pp. 730–737.
- M.D. Plumbley, Algorithms for non-negative independent component analysis, *IEEE Trans Neural Netw* 14 (2003), 534–543.
- S. Rickard, R. Balan, and J. Rosca, Real-time time-frequency based blind source separation, In *3rd Int Conf on Independent Component Analysis and Blind Source Separation (ICA2001)*, Dec 2001.
- S. Rickard and O. Yilmaz, On the approximate W-disjoint orthogonality of speech. In J. Principe and H. Bourlard (Editors), *Proc IEEE Int Conf on Acoustics, Speech and Signal Processing*, 2002, pp. 529–532.
- S.J. Roberts and R.M. Everson (Editors), *Independent components analysis: Principles and practice*, Cambridge University Press, Cambridge, UK, 2001.
- Z. Roth and Y. Baram, Multi-dimensional density shaping by sigmoids, *IEEE Trans Neural Netw* 7 (1996), 1291–1298.
- S.T. Roweis, One microphone source separation, In *Advances in neural information processing systems*, Vol. 13, MIT Press, Cambridge, MA, 2001, pp. 793–799.
- H. Sahlin and H. Broman, Separation of real-world signals, *Signal Process* 64 (1998), 103–113.
- J.M. Sanchis, F. Castells, and J.J. Rieta, Convolutional acoustic mixtures approximation to an instantaneous model using a stereo boundary microphone configuration, In *Fifth Int Conf on Independent Component Analysis*, Granada, Spain, Sept. 22–24, 2004, Lecture Notes in Computer Science, Vol. 3195, Springer-Verlag, pp. 816–823.
- P. Smaragdis, Blind separation of convolved mixtures in the frequency domain, *Neurocomputing* 22 (1998), 21–34.
- P. Smaragdis, Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs, In *Fifth Int Conf on Independent Component Analysis*, Granada, Spain, Sept. 22–24, 2004, Lecture Notes in Computer Science, Vol. 3195, Springer-Verlag, pp. 494–499.
- I. Takigawa, M. Kudo, A. Nakamura, and J. Toyama, On the minimum l_1 -norm signal recovery in underdetermined source separation, In *Fifth Int Conf on Independent Component Analysis*, Granada, Spain, Sept. 22–24, 2004, Lecture Notes in Computer Science, Vol. 3195, Springer-Verlag, pp. 193–200.

- A.C. Tang, B.A. Pearlmutter, M. Zibulevsky, and S.A. Carter, Blind separation of multichannel neuromagnetic responses, *Neurocomputing* 32–33 (2000), 1115–1120.
- F.J. Theis, A geometric algorithm for overcomplete linear ICA, In Workshop Report I of the Graduiertenkolleg, Windberg, 2001, pp. 67–76.
- K. Torkkola, Blind separation of convolved sources based on information maximization, In Proc IEEE Workshop on Neural Networks and Signal Processing (NNSP'96), Kyoto, Japan, 1996a, pp. 423–432.
- K. Torkkola, Blind separation of delayed sources based on information maximization, In Proc IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP'96), Atlanta, GA, 1996b, pp. 3509–3512.
- P.P. Vaidyanathan, *Multirate Systems and Filter Banks*. Prentice-Hall, Englewood Cliffs, NJ, 1993.
- M. van Hulle, Clustering approach to square and non-square blind source separation, In IEEE Workshop on Neural Networks for Signal Processing (NNSP), 1999, pp. 315–323.
- L. Vielva, D. Erdogmus, C. Pantaleon, I. Santamaria, J. Pereda, and J. Principe, Underdetermined blind source separation in a time-varying environment, In J. Principe and H. Bourslard (Editors), Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, Vol. 3, 2002, pp. 3049–3052.
- L. Vielva, D. Erdogmus, and J. Principe, Underdetermined blind source separation using a probabilistic source sparsity model, In 2nd Int Workshop on Independent Component Analysis and Blind Signal Separation, Helsinki, Finland, June 19–20, 2000, pp. 675–679.
- R. Vigário, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja, Independent component approach to the analysis of EEG and MEG recordings, *IEEE Trans Biomed Eng* 47 (2000), 589–593.
- T. Virtanen, Sound source separation using sparse coding with temporal continuity objective, In Proc Int Computer Music Conf (ICMC 2003), 2003.
- A. Westner and V.M. Bove, Blind separation of real world audio signals using overdetermined mixtures, In First Int Conf on Independent Component Analysis, 1999.
- B. Widrow, J. Glover, J. McCool, J. Kaunitz, C. Williams, R. Hearn, J. Zeidler, E. Dong, and R. Goodlin, Adaptive noise cancelling: Principles and applications, *Proc IEEE* 63(12) (1975), 1692–1716.
- N. Wiener, *Extrapolation, interpolation and smoothing of stationary time series*, MIT Press, Cambridge, MA, 1949.
- G. Wübbeler, A. Ziehe, B.-M. Mackert, K.-R. Müller, L. Trahms, and G. Curio, Independent component analysis of non-invasively recorded cortical magnetic DC-fields in humans, *IEEE Trans Biomed Eng* 47 (2000), 594–599.
- A. Yeredor, Approximate joint diagonalization using non-orthogonal matrices, *Proc ICA 2000*, June 2000, pp. 33–38.
- A. Yeredor, Blind source separation with pure delay mixtures, In Int Workshop on Independent Component Analysis and Blind Source Separation Conference (ICA), San Diego, CA, Dec. 9–12, 2001.
- O. Yilmaz and S. Rickard, Blind separation of speech mixtures via time-frequency masking, *IEEE Trans Signal Process* 52 (2004), 1830–1847.
- M. Zibulevsky, P. Kisilev, Y.Y. Zeevi, and B.A. Pearlmutter. Blind source separation via multinode sparse representation, In *Advances in neural information processing systems*, Vol. 14, MIT Press, Cambridge, MA, 2002, pp. 1049–1056.
- M. Zibulevsky and B.A. Pearlmutter, Blind source separation by sparse decomposition in a signal dictionary, *Neural Comput* 13 (2001), 863–882.
- A. Ziehe, K.-R. Müller, G. Nolte, B.-M. Mackert, and G. Curio, Artifact reduction in magnetoneurography based on time-delayed second order correlations, *IEEE Trans Biomed Eng* 47 (2000), 75–87.