Temporal Models in Blind Source Separation

Lucas C. Parra

Sarnoff Corporation, CN-5300, Princeton, NJ 08543

1 Introduction

Independent component analysis (ICA) aims to find statistical independent signals in an instantaneous linear mix. No knowledge of the original sources is assumed. It is therefore sometimes also referred to a Blind Source Separation (BSS). The concept of ICA was first introduced and formalized by Comon [8]. In resent years increasing interest in this concept arose in the neural network and signal processing community. There is wide range of possible applications. In the neural network community it has been interpreted as a version of the well know concept of redundancy reduction that has been repeatedly considered as an underlying principle in sensory formation [6, 19, 4, 10, 22]. For signal processing the interest arises in the context of sensor arrays, source reconstruction and location, various sonar applications, multidimensional blind channel equalization in multi-path coding, and more.

In the context of neural networks it can be formulated as finding a representation of the data with minimal mutual information among the output nodes of a network [10]. It can also be formulated as the representation that maximizes the information transmitted through a properly designed linear network [14, 7].

An explicit way of formulating this new principle is Maximum Likelihood [28]: assuming statistically independent model sources we try to find the model parameter that best explain the observations.

Maximum likelihood allows one to incorporate prior knowledge into the estimation procedure. Models of the temporal properties of the signals permit a sensible integration of the time coordinate to the statistical independence criteria. In principle this is easily accomplished by formulating the appropriate probability density function of the current model signals conditioned on their past [26]. This permits to incorporate signal models such as auto-regressive (AR) model into the ML approach, which improves the quality of the separation.

But time can be incorporated into the ML formulation in another very natural manner: instead of formulating the density function of an individual time sample we formulated the joint density of the signals within a time window, which allows the extension of the problem to the convolutive BSS case. In convolutive BSS the mixture arises as a combination of differently convolved independent source signals due to time delays and a reverberating acoustic environment.

The aim of this chapter is to discuss various optimization principles based on statistical independence focusing in particular to some aspects of time.

In section 2 we begin by relating the principles of minimum mutual information, and maximum transmitted information with statistical independence. In section 3 we will reformulate the problem of ICA, that is the recovery of source signals from the linear instantaneous mix, as a Maximum Likelihood problem. Next, in section 4, we will learn how to incorporate temporal context information into ML, leading to an improved ICA for instantaneous BSS. Finally we will see in section 5 how to extend the ML formulation to the case of convolutive BSS, while modeling the sources as an AR process. We focus there in particular to the application of separating multiple speakers in an reverberating environment, i.e. the signals arrive at the microphones with varying time delays and in differently convolved versions.

Note that inherently all these approaches capture higher order statistics of the signal, without which statistical independence cannot be obtained. However, in order to not divert the attention from the main focus of modeling temporal properties, we defer this important issue to an appendix.

2 Generating Independent Components

Assume we are given samples of random variables $(x_1, ..., x_N)^\top = \mathbf{x}$ distributed according to a probability density function $p(\mathbf{x})$. Furthermore consider a process that generates for a given $\mathbf{x}$ variables $(y_1, ..., y_N)^\top = \mathbf{y}$ distributed according to $p(\mathbf{y}|\mathbf{x};\mathbf{w})$. The transformation may be implemented by a (stochastic) neural network, where $\mathbf{w}$ is then the parameter vector of the network (see figure 1). The resulting output distribution is given by,

$$p(\mathbf{y}|\mathbf{w}) = \int d\mathbf{x} p(\mathbf{y}|\mathbf{x};\mathbf{w}) p(\mathbf{x}) \tag{1}$$

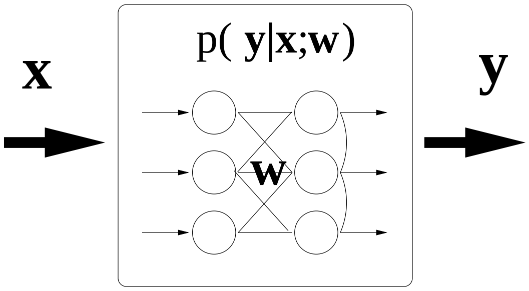

Fig. 1. Schematic representation of a network parametrized by $\mathbf{w}$ that should transform observations $\mathbf{x}$ into statistical independent variables $(y_1, ..., y_N)^\top = \mathbf{y}$

The purpose of this transformation is to obtain a new representation of $\mathbf{x}$ such that the new variables are statistical independent. Mathematically, statistical independence is expressed by the fact that the joint probability density of the variables $y_1, ..., y_N$ factors,

$$p(y_1, ..., y_N) = p(y_1)p(y_2)...p(y_N) = \prod_{i=1}^{N} p(y_i) \tag{2}$$

We will consider now different objective functions that measure how well the generated density (1) factors to produce independent components according to (2).

2.1 Minimal Mutual Information at the Output

An intuitive notion of independent variables is that they carry independent information. In other words, they carry minimal or no common information. According to Shannon the entropy $H[p(y)]$ of a probability density $p(y)$ captures how much information can be encoded by the random variable y,

$$H[p(y)] = -\int dyp(y) \ln p(y) \tag{3}$$

The information that is common to the variables y_i is measured by their mutual information,

$$MI[y_1; ...; y_N] = \sum_{i}^{N} H[p(y_i)] - H[p(\mathbf{y})] \tag{4}$$

The second term represents the joint entropy of the distribution, while the first term is the sum of the single coordinate entropies. Note that this expression is identical to the Kullback-Leibler distance (KLD) of the joint density (1) and the factorization (2),

$$KLD[p(\mathbf{y}), \prod_{i} p(y_i)] = \int d\mathbf{y}p(\mathbf{y}) \ln \left(\frac{p(\mathbf{y})}{\prod_{i}^{N} p(y_i)} \right) = \sum_{i}^{N} H[p(y_i)] - H[p(\mathbf{y})] \tag{5}$$

The KLD is a common distance measure between two distributions and captures here how well (1) factors. Mutual information will be therefore minimal, in fact zero, if the variables represent independent components.

Consider now a deterministic and invertible functional relation $\mathbf{y} = f(\mathbf{x}; \mathbf{w})$. We have then $p(\mathbf{y}|\mathbf{x}; \mathbf{w}) = \delta(\mathbf{y} - f(\mathbf{x}; \mathbf{w}))$ and (1) reduces to,

$$p_{\mathbf{y}}(\mathbf{y}|\mathbf{w}) = \left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right| p_{\mathbf{x}}(f^{-1}(\mathbf{y}; \mathbf{w})) \tag{6}$$

Taking the logarithm and the expectation over $p(\mathbf{x}, \mathbf{y})$ we obtain,

$$H[p(\mathbf{y}|\mathbf{w})] = H[p(\mathbf{x})] + E\left[\ln \left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right| \right] \tag{7}$$

If in addition the Jacobi determinant of the transformation is unity, $\left|\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right| = 1$, i.e. we have a volume conserving transformation, one can see that the information content of the input is equal to the information content of the output, i.e. $H[p(\mathbf{y}|\mathbf{w})] = H[p(\mathbf{x})]$. Since the entropy of the input density does not depend on the parameters $\mathbf{w}$, minimizing the mutual information (4) is in such a case equivalent to minimizing the entropy of the individual output coordinates. These considerations not only apply for linear but for any invertible non-linear transformation.

2.2 Maximum Transmitted Information

Surprisingly we find that under different conditions also *maximizing* the entropy of the output variables can lead to statistical independence. Consider the information that is common to the variables $\mathbf{x}$ and $\mathbf{y}$, that is, the information transmitted[1] through the mapping $\mathbf{x} \rightarrow \mathbf{y}$,

$$MI[p(\mathbf{x}), p(\mathbf{y})] = H[p(\mathbf{y})] - H[p(\mathbf{x}|\mathbf{y})] \tag{8}$$

The second term measures the randomness of the mapping. It has been argued [7] that for a deterministic mapping as discussed above, the second term can be ignored. Maximizing the transmitted information is therefore equivalent to maximizing the entropy of the output itself. Now, if every coordinate of the output is bounded by constants the maximum entropy will be given by a uniform distribution with, in fact, independent coordinates. In particular consider a linear transformation W with a bounded non-linearity $g(u)$ applied at each individual output (see figure 2),

$$\mathbf{y} = g(W\mathbf{x}) \tag{9}$$

In [21] it is shown more explicitly that if variables $\mathbf{x}$ were obtained from statistical independent coordinates $(s_1, ..., s_N)\top = \mathbf{s}$, distributed according to $p(\mathbf{s}_i)$, by a linear invertible transformation A,

$$\mathbf{x} = A\mathbf{s}\,, \tag{10}$$

then maximizing the transmitted information (8) with respect to W will converge towards $W^* = PDA^{-1}$. The matrices P and D are some appropriate permutation and diagonal scaling matrix, and do not change the fundamental result that W^* is an inversion of the mixing process A. These results holds, if the non-linearity was chosen such that it matches the source density according to $p(y) = \frac{\partial g(y)}{\partial y}$. Maximum entropy or maximum transmitted information is under these circumstances therefore equivalent to finding linear independent components.

[1] This expression is effectively the mutual information between input and output, and differs from the mutual information of the output coordinates discussed in the previous section.

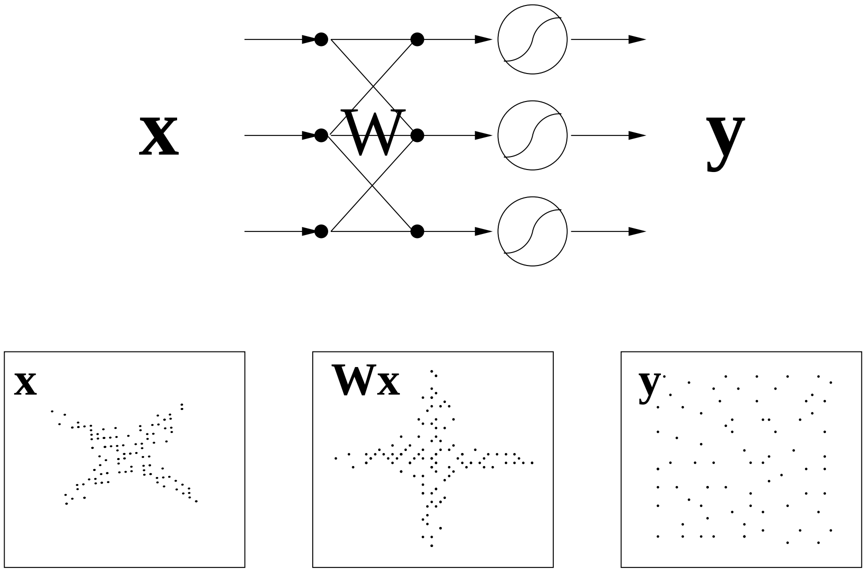

Fig. 2. Maximizing the information transmitted through this network (top) generates independent components at the output. At the bottom the distributions of a typical two dimensional case are depicted. If the non-linearity has been properly chosen, maximum transmitted information is equivalent to maximum entropy at the output. Its maximum in turn is for bounded non-linearities the uniform distribution (bottom, left) and is in fact statistical independent. It can be reached only if the output of the linear transformation $W\mathbf{x}$ is independent as well (bottom, center).

3 Finding Independent Components with Maximum Likelihood

The most explicit way, however, to express the independence assumption and to derive the optimization equations from such a transformation is the maximum likelihood (ML) criterion. The general idea of ML is to postulate a probability density for the observed variables that best describes them, eventually expressing how the observations where generated. The density will be parametrized by unknown properties, here for example the un-mixing coefficients W or parameters α describing the individual densities $p(s_i|\alpha)$ of the sources. One then tries to find the parameters that maximize the likelihood of the observations, i.e. the parameters that make the observation most likely.

Consider the linear un-mixing of (10),

$$\mathbf{y} = W\mathbf{x} \tag{11}$$

Note the slightly different definition of $\mathbf{y}$ here as opposed to (9). We are trying to explain the observations with a linear combination that has statistically independent coordinates. These do not correspond exactly to the original sources $\mathbf{s}$. They will only be the same up to permutation and scaling. Consider the likelihood of the mixtures $\mathbf{x}$, which depends on the model parameters W and α,

$$p_{\mathbf{x}}(\mathbf{x}|W,\alpha) = \left|\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right| p_{\mathbf{y}}(\mathbf{y}(W)|\alpha) = \left|\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right| \prod_{i}^{N} p_{y_i}(y_i|\alpha) = |W| \prod_{i}^{N} p_{y_i}(\mathbf{w}_i\mathbf{x}|\alpha) \quad (12)$$

Here $\mathbf{w}_i$ are the row vectors of the square matrix W. In order to find the right un-mixing parameters W we take the derivatives of the logarithm of (12), leading to,

$$\frac{\partial L(W,\alpha)}{\partial W} = \frac{\partial \ln p(\mathbf{x}|W,\alpha)}{\partial W} = W^{-\top} + \mathbf{u}\mathbf{x}^{\top} \quad (13)$$

with $\mathbf{u} = (\frac{\partial \ln p(y_1)}{\partial y_1}, ..., \frac{\partial \ln p(y_1)}{\partial y_1})^{\top}$. This gradient can be used to optimize W with stochastic gradient ascent. The inverse of W is however a expensive computation, and instead of taking the actual gradient we take its product with a positive definite matrix $W^{\top}W$. The resulting, so called natural gradient first introduced in [1] has a positive inner product with the original gradient, and points therefore into the same overall direction. This results now in the following update rules with a learning constant μ,

$$\mathit{\Delta} W = \mu\left(W + \mathbf{u}\mathbf{y}^{\top}W\right) \quad (14)$$

Note that the nonlinearity introduced in section 2.2 in a somewhat ad-hock manner emerges here in $\mathbf{u}$ very naturally. Also note that unlike the maximum entropy approach, where the densities of the individual sources must be known *a priori*, here we can also maximize the logarithmic likelihood with respect to parameters α, allowing for a flexible density function, that will try to match the individual source distributions.

4 Incorporating time into the ML source model

Before we turn to experimental results on separating multiple sound sources, we want to discuss how one can incorporate a better model of the temporal properties of sound into the single channel density $p(y|\alpha)$. We drop temporarily the index i. Consider conditioning the densities of the model sources by their past, i.e. $p(y(t)|y(t{-}1), y(t{-}2), ..., y(t{-}P)); \alpha)$. This allows us to model temporal relations of the signal. This concept has been called contextual ICA [26]. A standard signal processing model for temporal correlations of the signals is the linear auto-regressive (AR) model. The AR model makes a linear prediction $\bar{y}(t)$ of $y(t)$ from the past P samples,

$$e(t) = y(t) - \bar{y}(t) = y(t) - \sum_{\tau=1}^{P} a(\tau)y(t-\tau) \quad (15)$$

where $e(t)$ is considered to be the error of the prediction, and $a(\tau)$ the linear prediction coefficients (LPC). Recall that for the optimal LPC, i.e. the parameters that minimize the expected error $E[e(t)]$, the error signals are decorrelated in time [13]. The corresponding density function is then,

$$p(y(t)|y(t-1)...y(t-P);\mathbf{a}) = p(\mathbf{a}^\top \mathbf{y}(t)) \tag{16}$$

with $\mathbf{a} = (1, -a(1), .., -a(P))^\top$, and $\mathbf{y}(t) = (y(t), ..., y(t-P))^\top$. One can insert this density for every source into the likelihood function (12), where one may choose for every model source $y_i(t)$ independent AR parameters $\mathbf{a}_i$. The simplest approach for optimizing these parameters is again a stochastic gradient of the likelihood function $L(W, \mathbf{a}_1, ..., \mathbf{a}_N)$. The resulting update equations are,

$$\Delta a_i(\tau) = -\mu u_i(t) y_i(t-\tau), \text{ with } u_i(t) = \frac{\partial \ln p(y_i(t))}{\partial y_i(t)} \tag{17}$$

Figure 3 shows the separation results that were obtained for 10 different music sources, which were digitally mixed giving an instantaneous linear mix. A stationary AR model of size $P = 20$ was used, although good results were obtained also with $P = 5$. The density function for each channel was chosen as a zero mean Gaussian with unit variance. Gradient ascent rules (14) and (17) where used. The remaining cross-talk was hardly audible and corresponds to a signal-to-noise ration (SNR) between 10 and 100 for the 10 different channels.

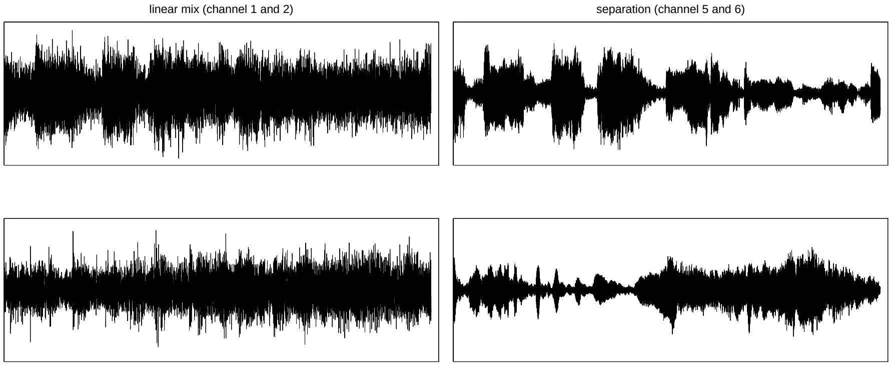

Fig. 3. left: Two of the 10 channels of 10 linearly mixed music CD sources. right: two channels of the output show good separation using contextual ICA. For more details see [27]

5 Convolutive Blind Source Separation and Modeling

In section 4 we have incorporated temporal models into the likelihood function in order to better model the individual sources. For instantaneous mixing in scenarios where the detectors are close, considering the speed of the signal in the medium, this is a sufficient way of considering time. Electrical signals with detectors in short distance and no substantial echoes is such a case. Acoustic sources however arrive at the different microphones in a typical reverberant environment with multiple echoes and different time delays. In such a case an instantaneous mix will not be sufficient. Essentially the signals at different time lags are mixed in different strengths. One has then to consider a convolutive mixture:

Assume N independent sources $(s_1(t)...s_N(t))^\top = \mathbf{s}(t)$ mixed in a unknown linear medium,

$$x_i(t) = \sum_{j=1}^{N} \sum_{\tau=0}^{\infty} h_{ij}(\tau) s_j(t-\tau) \tag{18}$$

We observe the mixtures $(x_1(t)...x_N(t))^\top = \mathbf{x}(t)$. To undo the effect of this causal filtering and mixing we require (eventually infinite size) non-causal finite impulse response (FIR) filters. We wish to find N statistically independent signals $(y_1(t)...y_N(t))^\top = \mathbf{y}(t)$ with a multidimensional, non-causal FIR filter $w(-K)...w(K)$ from the convolutive mixtures, where we limit us to a finite filter size K. Every $w(\tau)$ here represents a $N \times N$ unximing matrix for the time lag τ,

$$y_i(t) = \sum_{j=1}^{N} \sum_{\tau=-K}^{K} w_{ij}(\tau) x_j(t-\tau) \tag{19}$$

Note that we are not explicitly aiming to recover the original signals $\mathbf{s}(t)$ that lead to the mixtures $\mathbf{x}(t)$. We will merely try to model the mixtures by independent model sources $\mathbf{y}(t)$. The true sources $\mathbf{s}(t)$ may differ from these recovered independent signals $\mathbf{y}(t)$ by an arbitrary convolution and permutation [3]. We will try however to match the statistics of the sources by using linear prediction or signal subspace modeling techniques.

For the ML approach we require a density function of the observed signals as a function of the model parameters, which we will for now generically denote $\varPhi$. We will formulate the density function for a time window of the mixture signals $X(t) = (\mathbf{x}(t), ..., \mathbf{x}(t+T))$. This stands in contrast to previous formulations of the problem that have considered the likelihood of a single time instance only [26, 20, 3].

In order to express the density function in the space of the model sources we will consider the conditional density of the signals within the window condition on the signals outside the window which we will denote by $\tilde{X}(t) = \mathbf{x}(-\infty), ..., \mathbf{x}(t-1), \mathbf{x}(t+T+1), ..., \mathbf{x}(\infty)$.

$$p(X(t)|\tilde{X}(t);\varPhi) = \left|\frac{\partial Y(t)}{\partial X(t)}\right| p(Y(t)|\tilde{X}(t);\varPhi) \tag{20}$$

Here $Y(t) = (\mathbf{y}(t), ..., \mathbf{y}(t+T))$ is the corresponding window in the model source space[2]. The Jacobian $\frac{\partial Y(t)}{\partial X(t)}$ is a $NT \times NT$ matrix with coefficients,

$$\frac{\partial y_i(t)}{\partial x_l(r)} = \sum_{j=1}^{N} \sum_{\tau=-K}^{K} w_{ij}(\tau) \frac{\partial x_j(t-\tau)}{\partial x_l(r)} = w_{il}(t-r) \tag{21}$$

where $i, l = 0...N$ and $r = 0..K$, and $w(\tau)$ vanishes for values outside $-K \le \tau \le K$. It is useful to arrange the entries in X, Y such that the matrix $w(0)$ lies on the diagonal blocks,

$$\frac{\partial Y(t)}{\partial X(t)} = \begin{pmatrix} w(0) & w(-1) & ... & w(-T) \\ w(1) & w(0) & ... & w(1-T) \\ ... & ... & ... & ... \\ w(T) & w(T-1) & ... & w(0) \end{pmatrix} \equiv W \tag{22}$$

For a causal FIR the upper block triangle vanishes and the determinant in (20) is given by the determinant of $w(0)$,

$$\left|\frac{\partial Y(t)}{\partial X(t)}\right| = |w(0)|^T \ \text{ if } w(\tau) = 0 \text{ for } \tau < 0 \tag{23}$$

Although some have made this simplifying assumptions [5, 7], we wish to keep a non-causal filter, and will instead restrict ourself in section 5.2 to a circulant W in order to arrive to an efficient algorithm that can be implemented using the fast Fourier transform (FFT).

Now we introduce the independence assumption for the model sources by replacing in (20) the joint density of the model source by the product of the density of the individual sources,

$$p(X(t)|\tilde{X}(t);\varPhi) = |W(t)| \prod_{i=1}^{N} p(y_i(t)...y_i(t+T)|\tilde{X}(t);\varPhi) \tag{24}$$

[2] For the ML approach one requires the density of the observations $X(t)$ as a function of $X(t)$ itself and some model parameters $\varPhi$. Therefore, we have to replace $Y(t)$ in (20) by its definition (19). Note that is not possible to write $Y(t)$ as a function of $X(t)$ only. The model source values in the window at time t will depend by definition (19) on mixture values before and after the current frame. The conditioning of the probability on $\tilde{X}(t)$ is therefore a crucial step in order to make that substitution. In section 5.2 however, we consider periodic signals and the conditioning becomes superfluous.

5.1 Source Modeling

To our knowledge all current BSS algorithm make at this point for each ith model source a time independence assumption in (24) for the joint density of $y_i(t)...y_i(t+T)$ [16, 17, 20, 3, 5] This is for any reasonable acoustic signal not an appropriate model, and leads in their experiments to a whitened signal recovery.

The field of signal modeling in particular for speech enhancement offers a variety of ML approaches to single channel modeling. At this point many of these approaches can be combined to source separation by inserting the corresponding model probability into (24). We note however that all efficient algorithms are based on a linear dependency of the variables $(y(t)...y(t+T))^\top = \mathbf{y}(t)$ and mostly a Gaussian density[3].

For source separation however we require a non-Gaussian model since statistical independence is not uniquely defined for more than one Gaussian component in the mixture [33].

If one uses linear time correlation as described by a covariance matrix or linear prediction coefficients (LPC) the parameters introduced are equivalent to the parameters of the convolutions of the un-mixing FIR. The hope is however that the parameters describing the un-mixing and the parameters describing the source signal have different stationarity time scales. Speech for example will be stationary only within some 20ms - 40ms time frame, while the un-mixing coefficients should remain constant at least on a seconds scale, assuming that the location of the sources and the environment remains constant over that time. Single channel algorithms that adapt to varying statistics on a millisecond rate, as required by any single microphone speech enhancement algorithm, will extract to a certain extend the rapid varying portion of the linear correlations, while the slower converging source separation will pick up the slow varying time correlation due to the linear medium that mixes the source signals.

For short times on the order of 100-200 samples at 8kHz sampling rate the second order statistics of speech is well described with a multivariate Gaussian density. The covariance matrix however will change for larger time periods. The overall density will therefore be an accumulation or mixture of the instantaneous statistics. The net result of such a mixture is that the overall joint distribution will have high kurtosis, i.e. a strong mass at low amplitudes due to silence periods and long tales for high amplitude peeks. In the BSS literature the signal distribution has been therefore approximated by non-Gaussian distributions. The strongest approximation ignores time correlations, and assumes a high kurtosis one time step accumulated density $f(y)$,

$$p(y(t)...y(t+T)) \approx \prod_{\tau=0}^{T} f(y(t+\tau)) \tag{25}$$

A generalized Gaussian [17] has been used for $f(y)$ or a mixture of two zero

[3] We have ignored here and in the reminder of this section the model source index i since we are dealing with a individual channel. Bold notation now refers to the vector of signal values in the time window of size T for a single channel.

mean Gaussian with variances describing the silence and signal amplitudes [20]. A better approximation might result if we avoid the time independence assumption, and capture linear time correlations with a matrix $\Sigma(t)$ estimated in some window around t,

$$p(y(t)...y(t+T)) \approx f(\mathbf{y}^\top(t)\Sigma(t)\mathbf{y}(t)) \tag{26}$$

One might use also a multiple Gaussian model that allows for different covariance $\Sigma(t)$ for the silence, voiced and unvoiced, states up to a full hidden Markov model that incorporates state transition probabilities of the different sounds. These are routinely used for speech recognition resulting however in rather expensive models that require prior training.

We suggest a short term estimation of the linear correlations according the subspace or linear prediction methods used in speech enhancement [13, 18, 29, 12]. Assuming a correlation time P for which $\Sigma_{ij} = 0$ if $i - j > P$ we can expand the density of a frame of size T as,

$$p(y(t)...y(t+T)) = p(y(t)...y(t+P-1)) \prod_{\tau=t+P}^{t+T} p(y(\tau)|y(\tau-1)...y(\tau-P)) \tag{27}$$

The auto-regressive (AR) discussed in section 4 captures linear temporal correlations with the conditional density (16). Combining this source signal model for each of the N sources i with the source independence model we obtain the overall logarithmic likelihood[4],

$$\begin{aligned} L(W, \mathbf{a}_i...\mathbf{a}_N) = \ln|W| &+ \sum_{i=1}^{N} \sum_{\tau=t+P}^{t+T} \ln p(\mathbf{a}_i^\top \mathbf{y}_i(\tau)|\tilde{X}(\tau); W) \qquad (28) \\ &+ \sum_{i=1}^{N} \ln p(y_i(t)...y_i(t+P-1)|\tilde{X}(\tau); W) \end{aligned}$$

We will assume the LPC parameter to be constant within a time frame but change from frame to frame[5]. The un-mixing filters we assume constant throughout time. For $P << T$ we can neglect the last term here, and initialize the sum in the second term at $\tau = t$.

The extension done in this section compared to previous work is schematically depicted in figure 4. The present formulation should avoid the whitening of the model sources if the order P of the AR model is sufficiently large and the window size T in which to compute the AR parameters is sufficiently small to capture the fast variation of speech.

[4] The conditioning on $\tilde{X}(\tau)$ means nothing else that we can now substitute all $y()$ using (19). In the next section however it will be necessary to to assume periodic signals in order to obtain an efficient algorithm for updating W. The conditioning on the previous frame becomes then inconsequential.

[5] To be precise we should consider the likelihood of multiple frames by adding frame index k to $\mathbf{a}$, and setting t=k*T, while adding over all frames

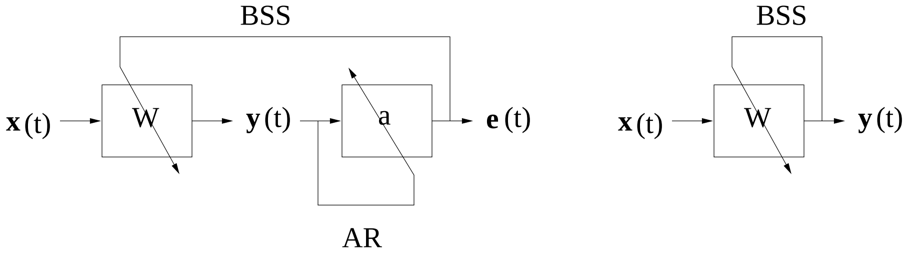

Fig. 4. Schematic representation of the suggested signal modeling and its relation to previous convolutive BSS algorithms. Left: The AR model whitens the signal producing the error signal $\mathbf{e}(t)$ which is used in turn for the BSS update. Details can be seen in the final update equation (42). Right: In previous work no modeling of temporal correlations of the model signal leads to separation and equalization.

If the densities can in fact be described in a short time frame of size P by a (zero mean) Gaussian, $\mathcal{N}(\mathbf{y}, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp(-\frac{1}{2}\, \mathbf{y}^\top \Sigma^{-1} \mathbf{y})$, we have,

$$p(y(t)...y(t+P)) = \mathcal{N}(\mathbf{y}(t), \Sigma) \tag{29}$$

The conditional density of a time sample $y(t)$ given its past P samples is then described by a one dimensional normal distribution of the error signal $e(t) = \mathbf{a}^\top \mathbf{y}(t)$ with a prediction accuracy σ,

$$p(y(t)|y(t-1)...y(t-P); \mathbf{a}, \sigma) = \mathcal{N}(\mathbf{a}^\top \mathbf{y}(t), \sigma) \tag{30}$$

where the parameters $\mathbf{a}$, σ are explicitly determined by the covariance matrix Σ,

$$\mathbf{a} = \Sigma_{+-} \Sigma_{-}^{-1} \text{ , and } \sigma = \Sigma / \Sigma_{-} = \Sigma_{+} - \Sigma_{+-} \Sigma_{-}^{-1} \Sigma_{-+} \tag{31}$$

$$\Sigma = \left(\begin{array}{c|c} \Sigma_{+} & \Sigma_{+-} \\ \hline \Sigma_{-+} & \Sigma_{-} \end{array} \right) = \left(\begin{array}{c|c} \sigma_{11} & \sigma_{12} \;...\; \sigma_{1P} \\ \hline \sigma_{21} & \sigma_{22} \;...\; \sigma_{2P} \\ ... & ... \;\; ... \;\; ... \\ \sigma_{P1} & \sigma_{P2} \;...\; \sigma_{PP} \end{array} \right) \tag{32}$$

We suggest to compute these coefficients from the sample autocorrelation matrix $\hat{\Sigma}$ estimated with the samples in the current window of size T. We may use the expressions above or more efficiently use Levinson-Durvin recursion on the signals $y(t)$ to compute the LPC from $\hat{\Sigma}$ [23]. This recursion gives an analytic solution for the LPC coefficients with a minimum least squares criterion, which is equivalent to the ML using the suggested signal model for a given frame [30].[6]

[6] To be precise the ML estimate of the LPC coefficients is computed with the so called covariance method which uses a somewhat different iteration [30].

5.2 Stochastic ML gradient

In other to optimize the logarithmic likelihood (28) also for the un-mixing filters W we use gradient descent. The main difficulty in deriving a gradient expression of (28) is to find a feasible expression for the derivative of the Jacobian (22). In fact, it will be necessary to assume that W is a circulant matrix,

$$W = (w_{ij}(n,m)) \equiv \begin{pmatrix} w(0) & w(T) & ... \, w(1) \\ w(1) & w(0) & ... \, w(2) \\ ... & ... & ... \; ... \\ w(T) & w(T-1) & ... \, w(0) \end{pmatrix} \tag{33}$$

The coefficients n, m denote the block column and row index, while the subscript index i, j refer to the index within each block. Columns are indexed therefore by i, n and rows by j, m. With this notation we can write now the gradient as,

$$\frac{\partial \ln |W|}{\partial w_{ij}(n,m)} = \sum_{n'm'}^{TT} \sum_{i'j'}^{NN} \frac{\partial \ln |W|}{\partial w_{i'j'}(n',m')} \frac{\partial w_{i'j'}(n',m')}{\partial w_{ij}(n,m)} = \sum_{n'm'}^{TT} \frac{|W_{ij}(n',m')|}{|W|} \delta^{n-m}_{n'-m'} \tag{34}$$

where $\delta^{z}_{z'} = 1$, if $modulo(z,(T+1)) = modulo(z',(T+1))$, and 0 otherwise. We have used the fact that for any invertible, square matrix A, $\frac{\partial \ln |A|}{\partial a_{ij}} = \frac{|A_{ij}|}{|A|}$, where $|A_{ij}|$ is the determinant of the matrix obtained after removing the ith row and jth column in A. Computing these determinants is an expensive operation of order $O(N^3 T^3)$. To avoid this we will use the argument commonly used in ICA algorithms, which was first introduced by Amari [1]. We multiply the gradients with a positive definite matrix $W^\top W$, to obtain the so called natural gradient. First consider

$$\begin{aligned} \left(\frac{\partial \ln |W|}{\partial W} W^\top \right)_{iu} (n,l) &= \sum_{mj}^{TN} \frac{\partial \ln |W|}{\partial w_{ij}(n,m)} w_{uj}(l,m) \\ &= \sum_{mj}^{TN} \frac{1}{|W|} \sum_{n'm'}^{TT} |W_{ij}(n',m')| \delta^{n-m}_{n'-m'} w_{uj}(l,m) \\ &= \frac{1}{|W|} \sum_{j}^{N} \sum_{n'm'}^{TT} |W_{ij}(n',m')| w_{uj}(l-n+n',m') \\ &= \frac{1}{|W|} \sum_{n'}^{T} \begin{cases} |W| \text{ if } n' = l-n+n' \text{ and } i = u \\ 0 \qquad\qquad \text{otherwise} \end{cases} \\ &= \frac{T}{|W|} I \end{aligned} \tag{35}$$

Read above $w_{ij}(n,m) = w_{ij}(modulo(n,(T+1)), modulo(m,(T+1)))$ if indexes n,m exceed their range $0..T$. Multiplying this identity matrix I with W finally leads to,

$$\frac{\partial \ln|W|}{\partial W} W^\top W = TW \tag{36}$$

Now we need to compute the gradient of the second term in (28).

$$\begin{aligned}
\frac{\partial \sum_{k\tau} \ln p(\mathbf{a}_k^\top \mathbf{y}_k(\tau))}{\partial w_{ij}(z)} &= \frac{\partial}{\partial w_{ij}(z)} \sum_{k=1}^{N} \sum_{\tau=t}^{t+T} \ln p(\sum_{\tau'=0}^{P} a_k(\tau') y_k(\tau-\tau')) & (37)\\
&= \sum_{\tau=t}^{t+T} g(\mathbf{a}_i^\top \mathbf{y}_i(\tau)) \sum_{\tau'=0}^{P} a_i(\tau') \frac{\partial}{\partial w_{ij}(z)} y_i(\tau-\tau') & (38)\\
&= \sum_{\tau=t}^{t+T} g(\mathbf{a}_i^\top \mathbf{y}_i^\top(\tau)) \sum_{\tau'=0}^{P} a_i(\tau') x_j(\tau-\tau'-z) & (39)
\end{aligned}$$

where $g(e) = \partial \ln p(e)/\partial e$. These coefficients for $z = 0..T$ represent the first column of the corresponding circulant matrix $\partial/\partial W$ arranged analogous to (33). In order to simplify the following multiplication with $W^\top W$ one has to assume periodic signals $\mathbf{x}(t)$, i.e. $\mathbf{x}(t) = \mathbf{x}(t+T+1)$. The model signals $\mathbf{y}(t)$ will then be periodic with period $T+1$ as well. This assumption not only simplifies the expressions but allows us to implement the convolutions with a discrete Fourier transform using a FFT. After some manipulations we obtain, again for the elements of the first column of a circulant matrix,

$$\left(\frac{\partial \sum_{k\tau} \ln p(...)}{\partial W} W^\top W\right)_{ij}(z) = \sum_{\tau=t}^{t+T} \sum_{u=1}^{N} \sum_{\tau'=t}^{t+T} g(\mathbf{a}_i^\top \mathbf{y}_i(\tau)) \mathbf{a}_i^\top \mathbf{y}_u(\tau+\tau'-z) w_{uj}(\tau') \tag{40}$$

These convolutions now can best be performed in the frequency domain. Transforming the natural gradient of (28), as expressed by the sum of (36) and (40), into the frequency domain involves applying the orthonormal coordinate transformation expressed by a matrix F with elements $F_{\nu\tau} = \frac{1}{\sqrt{T+1}} \exp(\frac{-\mathbf{i}\tau\nu 2\pi}{T+1})$, $\mathbf{i} = \sqrt{-1}$, which results for a circulant matrix like W into[7],

$$FWF^{-1} = diag(\mathcal{W}(0), ..., \mathcal{W}(T)) \tag{41}$$

That is, the Fourier coefficients $\mathcal{W} = F\mathbf{w}$ of the filter $\mathbf{w} = (w(0), ..., w(T))^\top$ represent the diagonal elements of a diagonal matrix. According to the convolution theorem the convolutions in (40) are performed by multiplying the Fourier coefficients independently. The overall gradients separate therefore in the frequency domain. Combining (36) and (40) and transforming the result into the

[7] Here we are writing for simplicity only the one-dimensional case. The multidimensional case is a trivial extension.

frequency domain we obtain the total natural gradient $\partial\mathcal{W}(\nu)$ in each frequency ν.

$$\partial\mathcal{W}_{ij}(\nu) = \mathcal{W}_{ij}(\nu) + G\left[\mathcal{A}_i(\nu)\mathcal{Y}_i(\nu)\right] \sum_{u=1}^{N} \mathcal{A}_i^*(\nu)\mathcal{Y}_u^*(\nu)\mathcal{W}_{uj}(\nu) \tag{42}$$

where $G[]$ is an operator that applies the function $g()$ in the temporal domain, $G[\mathcal{Y}] = Fg\left(F^{-1}Y\right)$, and the Fourier coefficients are given by, $\mathcal{A}_i = F(a_i(0), a_i(1), ..., a_i(P), 0, 0, ..., 0)^\top$ and $\mathcal{Y}_i = F\mathbf{y}_i$.

This result represents the extension of contextual ICA [26] to the convolutive case. It also represents a generalization of the equations suggested in [16, 17], which one obtains for $P = 1$, i.e. a time independence assumption. Note also that very recently [3] gave a explicit derivation of a natural gradient algorithm with infinite size FIR for the un-mixing leading to equations similar again to the ones proposed in [16, 17]. They do however not report results on real room recordings. In figure 5 we see the results obtain for two speakers in a noisy office environment. The separation improves the signal to background ration. It does however not separate the signal completely. The results depend on the type of signal and the choice of the density $p()$. This is a current subject of experimentation and study in the context of higher order statistics.

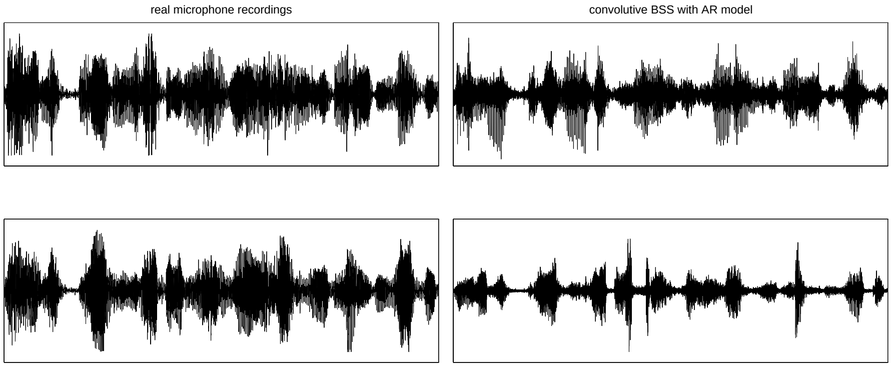

Fig. 5. Separation of real recordings with two microphones in a reverberant environment (office room) with algorithm (42) with $T = 512$, which corresponds at 8kHz to 64ms. The AR parameter were computed in each frame with the Levinson-Durvin recursion with $P = 20$.

At last one should mention that alternative convolutive BSS algorithms have been presented. In [31, 32] higher order statistics of the signals are explicitly measured in order to find separation filters. The issue of higher order statistics will be shortly outline in the appendix, section 7.

6 Summary

There are various ways to obtain separation of signals including minimal mutual information, maximum entropy, and maximum likelihood. Maximum likelihood in particular allows to incorporate time through additional signal models. It allows a rigorous extension of ICA to the convolutive case. Under the assumption of periodic signals one obtains then efficient update rules in the frequency domain.

Combining BSS with traditional single channel modeling techniques may allow for simultaneous BSS and signal enhancement leading to SNR improvements of the separated signals.

7 Appendix: ICA, PCA and Cumulants - Notes on higher order statistics

Statistical independence is inherently linked to the issue of higher order statistics. We have so far hardly mentioned higher orders statistics since we were including them in a implicit way. Either we used non-linear transformations that capture automatically more than second order of the signal as in section 2.2, or we have used with ML explicitly non-Gaussian distributions to model the signals capturing thus more than the second order. There is however a series of approaches and algorithms that explicitly include in their cost functions higher order statistics like moments, single coordinate cumulants, and cross-cumulants for multiple dimensions, which shall be discussed now briefly.

Interestingly principal component analysis (PCA) can be understood as a special case of statistical independence. In PCA one finds the rotation that minimizes the variance of the new coordinates $\mathbf{y}$. In fact, a rotation is a linear, volume conserving, and therefore entropy conserving transformation. As explained in section 2.1 minimal mutual information in $\mathbf{y}$ is in such a case equivalent to finding statistical independence. If we assume the variables $\mathbf{x}$ to be Gaussian distributed, then $\mathbf{y}$ will be Gaussian and minimizing mutual information at the output is equivalent to minimizing the variances of the output and is therefore equivalent to PCA [10, 25]. By making the Gaussian assumption we limit ourself to second order statistics. For a general input distribution however, second order statistic will not be sufficient to measure the entropies and obtain statistical independence [8, 33]. PCA will therefore not give statistical independence.

One approach that includes higher order statistics is to formulate a non-parametric model of the densities $p(y_i)$ of the individual outputs. Mostly it has been suggested to take expansions of the densities, like the Gram-Charlier, or Cramer-Edgeworth expansions [15]. The coefficients of these expansions are related to the higher order moments, or cumulants respectively. By doing so, one obtains expressions for single variable densities and the entropies as analytic functions of the higher order moments. By replacing the moments with their empirical estimates the entropy becomes essentially a function of the observations $\mathbf{y}$ and therefore a function of the parameters of the map. One can then use those expressions in a cost function base on any of the suggested cost-functions of section (2), and use gradient descent techniques to optimize the parameters of the input-output map [8, 24, 2].

An alternative approach of using higher order statistics is to formulate the conditions that *cross*-cumulants satisfy for statistical independent coordinates. The cross-cumulants are polynomial expressions of the cross-moments. Cross-moments of order q are defined by the expected values of all the possible combinations of powers $(q_1, ..., q_N) = \mathbf{q}$ with $q = \sum_i^N q_i$.

$$M[\mathbf{y}; \mathbf{q}] = \int d\mathbf{y} p(\mathbf{y}) y_1^{q_1} y_2^{q_2} ... y_N^{q_N} \tag{43}$$

Cumulants are essentially defined as the coefficients of the Taylor expansion of the logarithm of the Fourier transform of the density function about the zero frequency,

$$C[\mathbf{y}; \mathbf{q}] = \frac{\partial^q}{\mathbf{i}^q \partial^{q_1} \nu_1 ... \partial^{q_N} \nu_N} \ln \int d\mathbf{y} e^{\mathbf{i} \nu^\top \mathbf{y}} p(\mathbf{y}) \bigg|_{\nu = \mathbf{0}} \tag{44}$$

Cumulants $C[\mathbf{y}; \mathbf{q}]$ can be expressed entirely as specific polynomial combinations of the moments of the same or smaller order that use the same variables as selected by the particular $\mathbf{q}$ [9]. Cross-cumulants are important here since they can be shown to satisfy certain equations in the case of statistical independent variables [10]. Indeed, most cross-cumulants have to be zero. For example the elements of a covariance matrix represent the second order cross-cumulants. The off-diagonal terms vanish for statistical independent coordinates, expressing the fact that decorrelation is a necessary condition for statistical independence. While the third order cross-cumulants have to vanish as well, the fourth order cross-cumulants do not have to be all zero to guarantee independence [10]. One can combine those conditions in a single cost function. Again by replacing the cumulants with their sample estimates one obtains a cost that is a function of the parameters of the map. Minimizing that cost function with a gradient descent leads to independent components [10, 32].

Note the difference of the methods outlined above. While the second formulates conditions for the cross-cumulants the first approach tries to formulate a cost function in terms of single variable cumulants, i.e. diagonal terms of (44), and (43) with $q_1 = q$, or $q_2 = q$, ..., or $q_N = q$.

The criteria based on diagonal terms of cumulants have been used in instantaneous linear ICA [2] as well as in non-linear ICA [24, 25]. Cross-cumulants have been used in unsupervised learning of non-linear temporal recursion relations [11], as well as in convolutive ICA [31, 32] where cross-cumulants of coordinates at different time delays were considered.

While explicit consideration of higher order statistics tends to generate complicated and computationally expensive objective functions, they may converge faster and more reliably. The algorithms that include higher orders implicitly tend to simpler algorithms that are easier to implement efficiently. Their convergence properties may however be less favorable. The issue of which approach leads to the faster and more efficient algorithms may be settled as new and improved algorithms are developed.

References

1. S. Amari, A. Cichocki, and Yang A.A. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems 1995*, pages 752–763, Boston, MA, 1996. MIT Press.
2. S. Amari, A. Cichocki, and Yang A.A. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems 1995*, pages 752–763, Boston, MA, 1996. MIT Press.
3. S. Amari, S.C. Douglas, A. Cichocki, and A.A. Yang. Multichannel blind deconvolution using the natural gradient. In *Proc. 1st IEEE Workshop on Signal Processing App. Wireless Comm.*, Paris, France, 1997. IEEE.
4. J. Atick and A. Redlich. Towards a theory of early visual processing. *Neural Computation*, 2:308–320, 1990.
5. H. Attias and C.E. Schreiner. Blind source separation and deconvolution: Dynamic component analysis. *Neural Computations*, submitted, 1997.
6. H. Barlow. Sensory mechanism, the reduction of redundancy, and intelligence. In *National Physical Laboratory Symposium*, volume 10. Her Majesty's Stationery Office, London, 1959. The Mechanization of Thought Processes.
7. A. Bell and T. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.
8. P. Comon. Independent comonent analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.
9. Gardiner C.W. *Handbook of Stochastic Methods. Second Edition.* Springer-Verlag, New York, 1990.
10. G. Deco and Dragan Obradovic. *An Information Theoretic Approach to Neural Computing.* Perspective in Neural Computing. Springer, 1996.
11. G. Deco and B. Schuermann. Learning time series evolution by unsupervised extraction of correlations. *Physical Review E*, 41:1780–1790, 1995.
12. Yariv Ephraim and Harry Van Trees. A signal subspace approach fot speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 3(4):251–266, 1995.
13. H. Hayes, Monson. *Statistical Digital Signal Processing and Modeling.* Wiley, 1996.
14. C. Jutten and J. Herault. Blind separation of sources part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.
15. M.G. Kendal and A. Stuart. *The Advanced Theory of Statistics.* Charles Griffin & Company Limited, London, 1969.
16. R. Lambert and A. Bell. Blind separation of multiple speakers in a multipath environment. In *ICASSP 97*, pages 423–426. IEEE, 1997.
17. T. Lee, A. Bell, and R. Lambert. Blind separation of delayed and convolved sources. In *Advances in Neural Information Processing Systems 1996*, 1997.
18. J.S. Lim and A.V. Oppenheim. All-pole modelling of degraded speech. *IEEE Transaction ofn Acoustics, Speech, and Signal Processing*, 26(3):197–210, 1979.
19. R. Linsker. Self-organization in a perceptual network. *Computer*, pages 105–117, 21.

20. E. Moulines, J.F. Cardoso, and E. Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *ICASSP 97*, pages 3617–3620. IEEE, 1997.
21. J. Nadal and N. Parga. Non linear neurons in the low noise limit: a factorial code maximizes information transfere. *Network: Computation in Neural Systems*, 5(4):565–581, 1994.
22. B.A. Olhausen and D.J. Field. Emergence of simple-sell receptive field properties by learning sparce code for natural images. *Nature*, 381:607–609, 1996.
23. A.V. Oppenheim and R.W. Schafer. *Discret-Time Signal Processing.* Prentice Hall, 1989.
24. L. Parra. Symplectic nolinear component analysis. In *Advances in Neural Information Processing Systems 1995*, pages 437–443, Boston, MA, 1996. MIT Press.
25. L. Parra, G. Deco, and S. Miesbach. Redundancy reduction with information preserving nonlinear maps. *Network: Computation in Neural Systems*, 6:61–72, 1995.
26. B. Pearlmutter and L. Parra. A context-sencitive generalization of independent component analysis. In *International Conf. on Neural Information Processing*, Hong Kong, 1996.
27. B. Pearlmutter and L. Parra. Maximum likelihood source separation: a context-sensitive generalization of ica. In *NIPS 96*, Hong Kong, 1997.
28. D.T. Pham, Garrat P., and Jutten C. Separation of a mixture of independent sources through a maximum likelihood approach. In *Proc. EUSIPCO*, pages 771–774, 1992.
29. ML subspace signal enhancement. Fast adaptive eigenvalue decomposition: A maxmum likelihood approach. In *ICASSP 97*. IEEE, 1997.
30. Charles W. Therrien. *Discrete Random Signals and Statistical Signal Processing.* Prentice Hall, 1992.
31. E. Weinstein, M. Feder, and A.V. Oppenheim. Multi-channel signal separation by decorrelation. *IEEE Transaction on Spreech and Audio Processing*, 1(4):405–413, 1993.
32. D. Yellin and E. Weinstein. Multichannel signal separation: Methods and analysis. *IEEE Transaction on Signal Processing*, 44(1):106–118, 1996.
33. J. Zhu, X.R. Cao, and R.W. Liu. Blind source separation based on output independence - theory and implementation. In *NOLTA 95*, volume 1, pages 97–102, Tokyo, Japan, 1995. NTA Research Society of IEICE.