An Introduction to Independent Component Analysis and Blind Source Separation

Lucas C. Parra
Sarnoff Corporation, CN-5300, Princeton, NJ 08543
lparra@sarnoff.com

April 25, 1999

Contents

Abstract

Independent component analysis (ICA) and blind source separation (BSS) are related data analysis problems that have received considerable attention in the machine learning community during the last few years. This article gives a first year graduate level introduction to the subject. It treats the problem of finding independent components in an instantaneous linear mixture, and addresses the related subject of separating convolutive mixtures typical of an acoustic environment. After a straightforward derivation of a basic algorithm statistical independence is placed in a broader context of information principles such as redundancy reduction, minimum mutual information, and maximum information transmission. The relationship of statistical independence to higher order statistics, and entropy estimation is outlined. As we focus in particular on time series additional temporal statistics can be used to identify the unknown sources. To achieve separation in an ordered set, such as a time series, one can exploit non-stationarity and temporal correlations. This stays in contrast the common approach, which concentrates on higher order statistics of independent samples. This idea is carried over to the problem of convolutive mixtures, and a frequency domain algorithm is derived. It finds a multi-path filter that separate sources simultaneously recorded in a realistic acoustic environment.

1 ICA

1.1 Examples of linear mixtures of independent components

Independent component analysis (ICA) and blind source separation (BSS) refer to the problem of recovering statistically independent signals from a linear mixture. There is a variety of situations where we observe signals that originated as combination of independent processes or sources. Here are just a few examples:

- Cocktail-Party-Problem: Sound amplitudes in a acoustic environment add up linearly. Multiple sound sources such as speakers, music or noise sources are measured by the microphones as a mixture. The question is, how can one recover the individual speakers?
- Hyper-spectral sub-pixel identification: Hyper-spectral imagery consists of a set of images taken at different wavelengths - currently up to 200. Every material on the surface, i.e. rock, grass, trees, snow, etc. have different reflection coefficients at every wavelength. The area corresponding to a pixel contains usually a mixture of different surface materials, as the resolution is still in the range of a few square meters. The intensities in every pixel is therefore a linear

combination of the abundances of a materials and the reflection coefficients of each material. The problem is then, how to identify the surface materials and their abundance?

- Passive sonar: In passive sonar a large number of sensors (microphones) records signals originating form multiple sources such as shrimp noise, submarines engines, boats, etc. Every sensor records a different mixture as they are places in different locations and the amplitudes vary with distance form the sources. The task is to separate and identify the sound sources.

Effectively we find the problem of recovering sources from a linear mixture whenever there is independently generated signals, $\mathbf{s} = [s_1, ..., s_{d_s}]^\top$, a linear medium A, and a number of sensors to detect the mixtures $\mathbf{x} = [x_1, ..., x_{d_x}]^\top$, with $\mathbf{x} = A\mathbf{s}$. In the cocktail-party problem the time dependent sources $\mathbf{s}(t)$ correspond to d_s multiple speakers, while the mixtures parameters A correspond to the room response characteristic. In the hyper-spectral sub-pixel demixing problem the sources are the wavelength dependent reflectance $\mathbf{s}(\lambda)$ of the d_s different materials. The material abundances in the d_x different pixels represent the mixture coefficients A, while $\mathbf{x}(\lambda)$ represents the image at every spectral band λ.

For time sequences $\mathbf{s}(t)$ we call a mixture instantaneous if only signals from the immediate time t mix, i.e. $\mathbf{x}(t) = A\mathbf{s}(t)$. A convolutive mixture on the contrary may include time delays and echos $\mathbf{x}(t) = \sum_\tau A(\tau)\,\mathbf{s}(t)$. The hyper-spectral mixture explained above corresponds to the 'instantaneous' case since the wavelengths λ are not affected by the abundances, $\mathbf{x}(\lambda) = A\mathbf{s}(\lambda)$. Sound, however, travels slow compared to the distances of a typical acoustic environment, and the mixture is therefore convolutive. First the instantaneous mixture problem will be discussed. The convolutive case is treated in section 4.

1.2 Basic assumptions

Assume a *source* vector $\mathbf{s} \in R^{d_s}$, sampled according to density distribution $p(\mathbf{s})$, where the coordinates are statistical independent, i.e.

$$p(s_1, ..., s_{d_s}) = p(s_1)p(s_2)...p(s_{d_s}). \tag{1}$$

The sources $\mathbf{s}$ are not observed directly nor is the particular form of the individual distributions $p(s_i)$ known, hence the name *blind* source separation. Further assume the coordinates are "mixed" linearly:

$$\mathbf{x} = A\mathbf{s} \tag{2}$$

and we observe only the mixture $\mathbf{x} \in R^{d_x}$

Task: Given a set of such $\mathbf{x}$ find A and therefore the original independent sources $\mathbf{s}$.

Note that any permutation and scaling of independent variables remains independent. In fact, in the following sections the recovered model sources are often referred to as coordinates $\mathbf{y}$, generated by some linear inversion, $\mathbf{y} = W\mathbf{x}$. Statistical independence specifies the model sources $\mathbf{y}$ only up to an arbitrary scaling, expressed here as a diagonal matrix D, and permutation P, i.e. $\mathbf{s} = PD\mathbf{y}$. For simplicity this will be ignored in the notation at times, and we identify the estimated sources as $\hat{\mathbf{s}} = \mathbf{y}$.

1.3 ICA from Maximum Likelihood

Whenever something is known about the probability of the generating process one should think of using Maximum Likelihood (ML), as it explicitly utilizes that knowledge in the parameter estimation[1]). In ML one needs to express the probability of the data for given parameters in terms of the assumed model distribution. In the present case we need

$$p_x(x|A) \text{ in terms of } p_s(s)$$

Recall that a change of variable in a density will involve the Jacobi determinant of the transformation

$$p_x(x) = |ds/dx| p_s(s) \tag{3}$$

In terms of the inverse $A^{-1} \equiv W = [\mathbf{w}_1, ..., \mathbf{w}_{d_s}]^\top$ and using (1) this leads to

$$p_{\mathbf{x}}(\mathbf{x}|W) = |W|\, p_{\mathbf{s}}(W\mathbf{x}) = |W| \prod_{i=1}^{d_s} p_{s_i}(\mathbf{w}_i^\top \mathbf{x}) \tag{4}$$

Consider a set of N independent, and identically distributed (i.i.d.) such observations $\mathbf{x}_1, ..., \mathbf{x}_N$. The likelihood of observing the entire dataset is then

$$p(\mathbf{x}_1, ..., \mathbf{x}_N|W) = |W|^N \prod_{j=1}^{N} \prod_{i=1}^{d_s} p_{s_i}(\mathbf{w}_i^\top \mathbf{x}_j) \tag{5}$$

The ML estimate of A or equivalently its inverse W is now given by

$$W_{ML} = \arg\max_W \ln p(\mathbf{x}_1, ..., \mathbf{x}_N|W) \tag{6}$$

To maximize the log-likelihood one can use stochastic gradient ascent. This means we apply the gradient with respect to W of the log-likelihood of every sample j,

[1] See [9] for Maximum Likelihood

$$\frac{\partial \ln p(\mathbf{x}_j|W)}{\partial W} = W^{-\top} + \mathbf{u}_j \mathbf{x}_j^\top \tag{7}$$

with $\mathbf{u} = [\partial \ln p(s_1)/\partial s_1, ..., \partial \ln p(s_{d_s})/\partial s_{d_s}]^\top$. Note that obtaining the inverse W^{-1} at every sample is computationally intensive. Instead of taking the actual gradient, however, one can 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,

$$\Delta W = \mu \left(W + \mathbf{u}\mathbf{s}^\top W \right) \tag{8}$$

As the individual source distributions $p(s)$ are unknown, one can represent them with an appropriate parametric model $p(s|\alpha)$, and obtain a likelihood $p(x|W, \alpha)$ with additional parameters α. We find then the optimal parameter using additional gradients of the likelihood with respect to α. In figure 4 of section 3.1 a mixture model[2] is used, and the mixture parameters are determined in the same ML formalism.

1.4 PCA and ICA

It is interesting to note that the well known Principal Component Analysis (PCA) represents a special case of ICA under the two following constraints

- W are restricted to rotations, i.e $W^{-1} = W^\top$
- s is Gaussian distributed, i.e. $p(s) = (2\pi\lambda)^{-1/2} \exp(-s^2/(2\lambda))$

To see this consider the ML solution (6). At the solution the sum off the gradients (7) for all data must vanish,

$$0 = NW^{-\top} + \sum_{j=1}^{N} \mathbf{u}_j \mathbf{x}_j^\top, \text{ or equivalently } W^{-\top} = -\frac{1}{N} \sum_{j=1}^{N} \mathbf{u}_j \mathbf{x}_j^\top \tag{9}$$

For a Gaussian distribution we have $u = \partial \ln p(s)/\partial s = -s/\lambda$, and using the notation $\Lambda = \text{diag}(\lambda_1, ..., \lambda_{d_s})$ one can write

$$W^{-\top} = \frac{\Lambda^{-1}}{N} \sum_{j=1}^{N} \mathbf{s}_j \mathbf{x}_j^\top = \frac{\Lambda^{-1}}{N} \sum_{j=1}^{N} W \mathbf{x}_j \mathbf{x}_j^\top = \Lambda^{-1} W R_x \tag{10}$$

Further using $W^{-1} = W^\top$ we obtain,

$$R_x = W \Lambda W^\top \tag{11}$$

The solutions of (11) are the rotations W that diagonalize the covariance matrix, which is one definition of PCA.

[2] see [9] for mixture models.

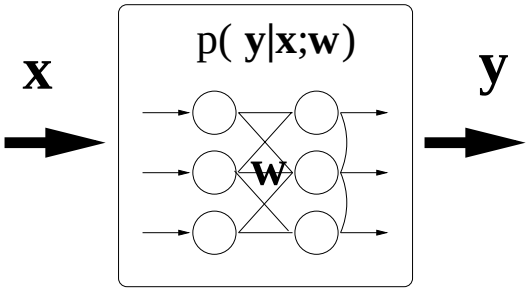

Figure 1: Schematic representation of a network parametrized by $\mathbf{w}$ that should transform observations $\mathbf{x}$ into statistical independent variables $(y_1, ..., y_N)^\top = \mathbf{y}$

1.5 Minimal Mutual Information

The ML approach treat ICA as a fairly straight forward parameter estimation problem. In this and the following section we will take a more general view of statistical independence base on information theoretic principles.

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,

$$p(\mathbf{y}|\mathbf{w}) = \int d\mathbf{x} p(\mathbf{y}|\mathbf{x}; \mathbf{w}) p(\mathbf{x}) \tag{12}$$

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 $y_1, ..., y_{d_y}$ factors,

$$p(y_1, ..., y_{d_y}) = p(y_1)p(y_2)...p(y_{d_y}) = \prod_{i=1}^{d_y} p(y_i) \tag{13}$$

We will consider now different objective functions that measure how well the generated density (12) factors to produce independent components according to (13).

An intuitive notion of independent variables is that they carry independent information. In other words, they carry minimal or no common or mutual information (MMI). 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,

$$H[p(y)] = -\int dy p(y) \ln p(y) \tag{14}$$

The information that is common to the variables y_i is measured by their mutual information [5],

$$MI[y_1; ...; y_{d_Y}] = \sum_{i=1}^{d_y} H[p(y_i)] - H[p(\mathbf{y})] \tag{15}$$

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 (12) and the factorization (13),

$$\begin{aligned} KLD[p(\mathbf{y}), \prod_i p(y_i)] &= \int d\mathbf{y} p(\mathbf{y}) \ln \left(\frac{p(\mathbf{y})}{\prod_{i=1}^{d_y} p(y_i)} \right) \\ &= \sum_{i=1}^{d_y} H[p(y_i)] - H[p(\mathbf{y})] \end{aligned} \tag{16}$$

The KLD is a common distance measure between two distributions [5], and captures here how well (12) 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 (12) reduces to,

$$p_{\mathbf{y}}(\mathbf{y}|\mathbf{w}) = \left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right| p_{\mathbf{x}}(f^{-1}(\mathbf{y}; \mathbf{w})) \tag{17}$$

Taking the logarithm and the expectation over $p(\mathbf{x}, \mathbf{y})$, denoted as $\langle ... \rangle$, we obtain,

$$H[p(\mathbf{y}|\mathbf{w})] = H[p(\mathbf{x})] + \left\langle \ln \left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right| \right\rangle \tag{18}$$

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 (15) 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.

1.6 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[3] transmitted through the mapping $\mathbf{x} \rightarrow \mathbf{y}$,

$$MI[p(\mathbf{x}), p(\mathbf{y})] = H[p(\mathbf{y})] - H[p(\mathbf{x}|\mathbf{y})] \tag{19}$$

The second term measures the randomness of the mapping. It has been argued [3] 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, and is sometimes referred to as the InfoMax principle. 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),

$$\mathbf{y} = g(W\mathbf{x}) \tag{20}$$

In [15] 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,

$$\mathbf{x} = A\mathbf{s}\,, \tag{21}$$

then maximizing the transmitted information (19) 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 result holds only if the non-linearity 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.

2 Higher order statistics, and entropy estimation

2.1 Statistical independence and higher order statistics

Statistical independence is inherently linked to the issue of higher order statistics. The relation can be expressed fairly directly. For two random

[3] 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.

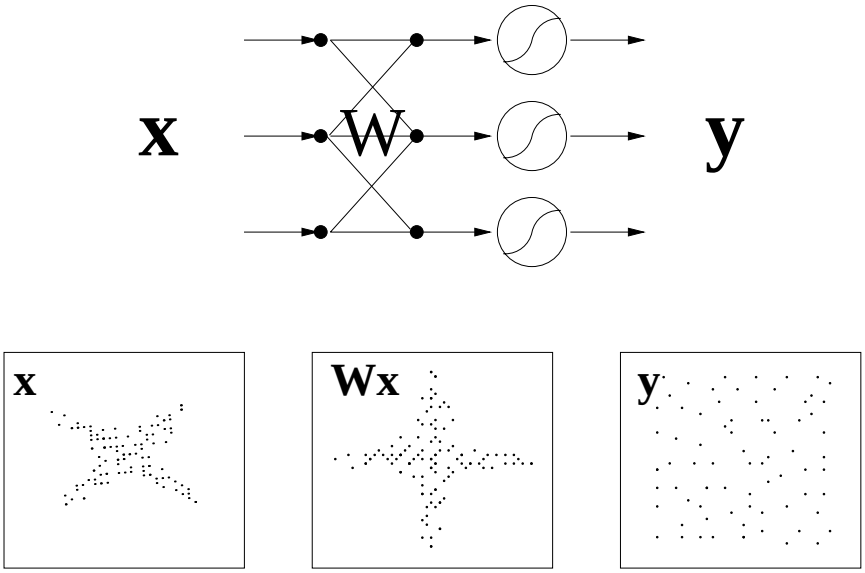

Figure 2: Maximizing the information transmitted through this network (top) generates independent components at the output. At the bottom the distributions of a typical two dimensional case are depicted. If the non-linearity has been properly chosen, maximum transmitted information is equivalent to maximum 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).

variables a, and b

$$p(a,b) = p(a)p(b) \leftrightarrow \langle f(a)g(b)\rangle = \langle f(a)\rangle \langle g(b)\rangle \tag{22}$$

where $\langle ...\rangle$ corresponds to the expectation over the random variables. In particular the higher moments of statistical independent variables satisfy

$$\langle a^n b^m\rangle = \langle a^n\rangle \langle b^m\rangle \text{ for all } n, m \tag{23}$$

Therefore, given a good estimate of all higher moments and cross-moments, one can verify if the variables are statistical independent. By optimizing the linear transformation W, such that all moments factor, one can obtain independent components. In practice in such an approach one has to restrict to a limited number of moments. By doing so, one is effectively making an assumption as of which moments are important to represent the density $p(a)$ and $p(b)$.

In the previous sections we could ignore higher order statistics since they were included implicitly through the derivative of the model source densities $\partial p(s)/\partial s$ or equivalently through non-linearities $g()$.

In the case of **known** source densities we saw that with the ML and the InfoMax approach we may limit ourself to $f(a) = a$. For that case much work has been done to show which choice of $g()$ or equivalently the derivative of $p(b)$ is appropriate for a particular class of underlying distributions.

In the case of **unknown** source densities one can try to approximated the distributions by using the higher order statistics explicitly. With the MMI approach described before one can limit oneself to measuring *single coordinate* statistics. In that case the entropy of the model variables y_i has to be estimated and minimized. This will be outlined in the following section. Alternatively one can also try to find model sources that satisfy *multi-coordinate* conditions of the sort of (23) among the different model sources. This involves cross-moments or cross-cumulants, which will be discussed in section 2.4.

2.2 Entropy estimation and ICA

In a sense all higher order techniques represent different approximations of the density function of the model sources $p(y)$ required in any of the approaches discussed so far (ML, minimal mutual information, or InfoMax). They all relate to the fundamental problem of how to estimate the entropy $H[p(y)]$ of a continuous variable, given a limited number of observations $y_1, y_2, ..., y_N$. The entropy to be optimized is usually approximated as,

$$H[p(y)] = \int dy\, p(y) \ln p(y) \approx \sum_{j=1}^{N} \ln \hat{p}(y_j) \tag{24}$$

where $\hat{p}(y)$ represents some estimate of $p(y)$.

The subject of density estimation goes far beyond the scope of this presentation (see [9]). But in principle one divides between parametric and non-parametric techniques for estimating a density.

A *parametric* technique defines a family of density functions $p(y|\alpha)$ in terms of a set of parameters α. The parameters are then optimized, e.g. with ML, so that the density function corresponds to the observed samples. In the context of ICA different parametric representations $p(y|\alpha)$ have been used. Examples include mixture models, as in the example of figure 4, and generalized Gaussian.

The *non-parametric* techniques usually define the estimated density directly in terms of the observed samples $y_1, ..., y_N$, i.e. $p(y|y_1, ..., y_N)$. For small datasets, i.e. small N, they are often faster and simpler to obtain as they do not require any adaption or training process. The best known non-parametric estimate is the histogram, which is fairly data intensive. Somewhat less data is required by the Parzen-Windows method [9]. Other non-parametric techniques are based on higher moments [13], e.g. Gram-Charlier expansion, Parson densities, or on higher cumulants [13], e.g. Edgeworth expansion. These techniques require less data as they make some general assumptions about the distributions, e.g. a single central mode, symmetry, Gaussian tails, and the like.

2.3 Entropy estimation with cumulants

Here we will demonstrate the use of cumulants for entropy estimation, and optimization with the objective of finding independent components.

In a seminal paper, effectively defining the term ICA, Comon [4] suggests to use the Edgeworth expansion of a probability distribution. This expansion is an analytic expression of the entropy in terms of measurable higher order cumulants. Edgeworth expands the multiplicative correction to the best Gaussian approximation of the distribution in the orthonormal basis of Hermite polynomials $h_\alpha(y)$. The expansion coefficients are basically given by the cumulants c_α of the distribution p(y).[4] The Edgeworth expansions reads for a zero-mean distribution with variance σ^2, see [13],

$$\begin{aligned} p(y) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}} f(y) \\ f(y) &= 1 + \frac{c_3}{6\sigma^3} h_3(\frac{y}{\sigma}) + \frac{c_4}{24\sigma^4} h_4(\frac{y}{\sigma}) + \frac{c_5}{120\sigma^5} h_5(\frac{y}{\sigma}) + ... \end{aligned} \tag{25}$$

Note, that by truncating this expansion at a some order, we obtain an approximation, which may not be strictly positive. Figure 3 shows a sampled exponential distribution with additive Gaussian noise.

[4] Cumulants c_α can be expressed in terms of moments m_α. The first five cumulants for a zero mean distribution ($c_1 = m_1 = 0$) are given by: $c_2 = m_2, c_3 = m_3, c_4 = m_4 - 3m_2^2, c_5 = m_5 - 10m_3m_2$. Moments can be estimated with the sampled data points $\{y_1, .., y_N\}$: $\hat{m}_\alpha = \frac{1}{N}\sum_{i=1}^{N} y_i^\alpha$

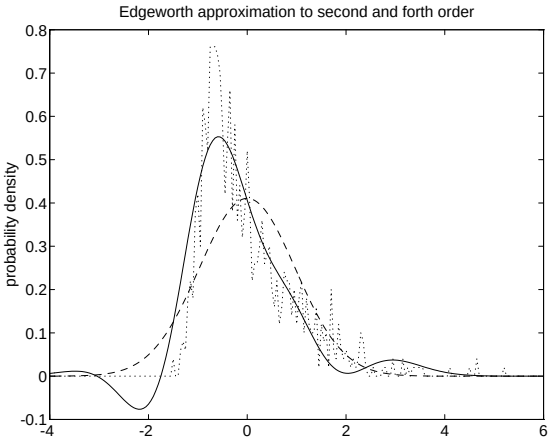

Figure 3: Doted line: exponential distribution with additive Gaussian noise sampled with 1000 data points. (noise-variance/decay-constant = 0.2). Dashed line: Gaussian approximation equivalent to the Edgeworth approximation to second order. Solid line: Edgeworth approximation including terms up to fourth order.

The coefficients of these expansions are related to the higher order moments, or cumulants respectively. The first cumulant c_1 is the *mean* of a distribution. The second cumulant c_2 represents the *variance*. For symmetric distributions the third order cumulant c_3 vanishes - it is therefore sometimes also referred to as *skew*. The fourth order cumulant is commonly referred to as *kurtosis*, $k = c_4/c_2^2$. Kurtosis represents how peeked the mode, or how long the tails of a distribution are. From (25) we see that all cumulants higher than the second must vanish for a Gaussian distribution.

By using this approximation in terms of moments and cumulants one obtains expressions for single variable densities and the entropies as analytic functions of the higher order moments, i.e. $p(y) = p(y|c_1, c_2, ...)$ or $H[p(y)] = H(c_1, c_2, ...)$. The cumulants can be estimated using the the samples $y_1, ..., y_N$, which in turn are a function of the mixture parameters, i.e. $c_i \approx \hat{c}_i(y_1, ..., y_N) = \hat{c}_i(W)$. Combining all this one obtains expressions of the entropy in terms of the mixture parameter up to a desired order n,

$$\begin{aligned} H[p(y)] &= \frac{1}{2}\ln(2\pi e) + \frac{1}{2}\ln\sigma^2 - \frac{1}{12}\frac{c_3^2}{\sigma^6} - \frac{1}{48}\frac{c_4^2}{\sigma^8} - \frac{7}{48}\frac{c_3^4}{\sigma^{12}} + \frac{1}{8}\frac{c_3^2}{\sigma^6}\frac{c_4}{\sigma^4}... \\ &\approx H(\hat{c}_1(W), \hat{c}_2(W), ..., \hat{c}_n(W)) \end{aligned} \tag{26}$$

This can be used in the cost function (15) of the MMI approach, and

the solution is found with a gradient descent techniques to optimize the parameters W [4, 17, 1].

One can show that cutting off after the second order term in (26) gives an upper bound on the entropy, $H[p(y)] \leq \frac{1}{2}\ln\sigma^2 + const.$, which can be used as an alternative and simplified cost function [18, 7]. In particular for a Gaussian input distribution and a linear transformation the upper bound becomes an strict equality. The MMI criterion, i.e. minimizing (15), reduces then to finding the minimum of the sum of the log of the output variances $\sigma_i^2 \approx \hat{\sigma}_i^2(W)$ under the constraint that entropy is conserved,

$$W_{MMI} = \arg\min_{W,|W|=1} \sum_{i=1}^{d_y} \ln \hat{\sigma}_i^2(W) \tag{27}$$

Again one can show that for rotations, $W^{-1} = W^\top$, this reduces to PCA [7] [5]

2.4 Cross-moments and Cross-cumulants

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 $(q_1, ..., q_{d_y}) = \mathbf{q}$ with $q = \sum_{i=1}^{d_y} q_i$.

$$M[\mathbf{y};\mathbf{q}] = \int d\mathbf{y} p(\mathbf{y}) y_1^{q_1} y_2^{q_2} ... y_{d_y}^{q_{d_y}} \tag{28}$$

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,

$$C[\mathbf{y};\mathbf{q}] = \frac{\partial^q}{\mathbf{i}^q \partial^{q_1}\nu_1 ... \partial^{q_{d_y}}\nu_{d_y}} \ln \int d\mathbf{y} e^{\mathbf{i}\nu^\top \mathbf{y}} p(\mathbf{y}) \Big|_{\nu=\mathbf{0}} \tag{29}$$

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 $\mathbf{q}$ [6]. Cross-cumulants are important here since they can be shown to satisfy certain equations in the

[5] To show this one can use the fact that for a Gaussian input distribution $p(\mathbf{x})$ with covariance R_x and a linear, square, and invertible transformation W the variances of the output, $\mathbf{y} = W\mathbf{x}$, can be written as $[\sigma_1^2, ..., \sigma_{d_y}^2] = diag(WR_xW^\top)$. To enforce the constraint $|W| = 1$ one can rescale the arbitrary W to $\tilde{W} = W/|W|^{1/d_y}$. With that, the variances become $diag(WR_xW^\top/|W|^{2/d_y})$. Inserting this into the criteria (27) and setting the derivative with respect to W to zero leads to the now familiar diagonalization equations. In the case of rotations this corresponds the eigenvalue problem of PCA.

case of statistical independent variables [7]. 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 [7]. 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 [7, 25].

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 (29), and (28) 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 [1] as well as in non-linear ICA [17, 18]. Cross-cumulants have been used in unsupervised learning of non-linear temporal recursion relations [8] (see section 3.3), as well as in convolutive ICA [24, 25] 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 depend less on the accuracy of the assumptions on the source densities. The algorithms that implicitly include higher orders tend to simpler algorithms that are easier to implement efficiently.

3 Source separation on time series

This far we have ignored the fact that the signals are often observed in a particular ordering $\mathbf{x}(t), \mathbf{x}(t+1), \mathbf{x}(t+2), ...$, which may containing useful information. The most immediate way of measuring temporal statistic is the auto-correlation and cross-correlations, which will be used in the next two sections. In fact, section 3.2 shows that for non-white or non-stationary signals one can recover the unknown mixture coefficients A using only second order statistic, and without any additional assumptions or knowledge on the source distributions $p(y)$. But first, we will see how to include temporal structure in the ML estimation approach outlined in section 1.3.

3.1 ICA and AR model of the signal

In section 1.3 we assumed that multiple samples $\mathbf{x}(t), \mathbf{x}(t+1), \mathbf{x}(t+2), ...$ are identically and independently distributed according to some $p(\mathbf{x})$. For time series this is often not the case. Instead, subsequent samples are correlated. In terms of probabilities this can be expressed by conditional densities $p(x(t)|x(t-1), x(t-2), ..., x(t-P))$, where the current sample $x(t)$

depends on the its past values. The joint distribution of the observations should be factored, rather than as a simple product, as in (5), into a product of conditional densities

$$\begin{aligned} & p(\mathbf{x}(t),\mathbf{x}(t-1),\mathbf{x}(t-2),...) \\ & \quad = p(\mathbf{x}(t)|\mathbf{x}(t-1),\mathbf{x}(t-2),...)\ p(\mathbf{x}(t-1),\mathbf{x}(t-2),...) \qquad (30) \\ & \quad = ... = \prod_{\tau=0}^{\infty} p(\mathbf{x}(t-\tau)|\mathbf{x}(t-\tau-1),\mathbf{x}(t-\tau-2),...) \qquad (31) \end{aligned}$$

This allows us to model temporal relations of the signal. In section 1.3 we expressed the likelihood of observations $p(\mathbf{x}|W)$ in terms of the single variable densities $p_{s_i}(s_i)$. Now the conditional density $p(\mathbf{x}(t)|\mathbf{x}(t-1),\mathbf{x}(t-2),...;W)$ will be expressed in terms of $p(s_i(t)|s_i(t-1),s_i(t-2),...)$ in an analogous way.

$$\begin{aligned} & p(\mathbf{x}(t),\mathbf{x}(t-1),\mathbf{x}(t-2),...) \qquad (32) \\ & \quad = \prod_{\tau=0}^{\infty} p(\mathbf{x}(t-\tau)|\mathbf{x}(t-\tau-1),\mathbf{x}(t-\tau-2),...) \qquad (33) \\ & \quad = \prod_{\tau=0}^{\infty} |W| \prod_{i=1}^{d_s} p(s_i(t-\tau)|s_i(t-\tau-1),s_i(t-\tau-2),...)\Big|_{\mathbf{s}=W\mathbf{x}} \qquad (34) \end{aligned}$$

This concept has been called contextual ICA [20]. A standard signal processing model for temporal correlations of the signals is the linear autoregressive (AR) model. The AR model makes a linear prediction $\bar{s}(t)$ of $s(t)$ from the past P samples,

$$e(t) = s(t) - \bar{s}(t) = s(t) - \sum_{\tau=1}^{P} a(\tau)s(t-\tau) \qquad (35)$$

where $e(t)$ is considered to be the error of the prediction, and $a(\tau)$ the linear prediction coefficients (LPC). For a Gaussian distributed error the ML estimate of the LPC are given by the parameters that minimize the expected error $\left\langle e^2(t)\right\rangle$. Recall that for those the error signals are decorrelated in time [10]. For arbitrarily distributed error the corresponding density function is then,

$$p(s(t)|s(t-1)...s(t-P);\mathbf{a}) = p(\mathbf{a}^\top \mathbf{s}(t)) = p(e(t)) \qquad (36)$$

with $\mathbf{a} = [1,-a(1),..,-a(P)]^\top$, and $\mathbf{s}(t) = [s(t),...,s(t-P)]^\top$. One can insert this density for every source into the likelihood function (34),

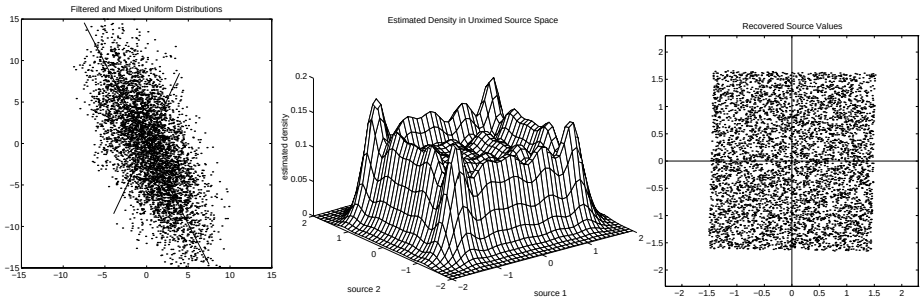

Figure 4: Left: samples of two independent, uniformly distributed variables after low-pass filter and linear mix. Center: density estimated for residuals e_i. The resulting density is a product of single coordinate Gaussian mixture models. Right: recovered independent model sources.

where one may choose for every model source $s_i(t)$ independent AR parameters $\mathbf{a}_i$. The simplest approach for optimizing these parameters is again a stochastic gradient of the likelihood function $L(W, \mathbf{a}_1, ..., \mathbf{a}_{d_s}) = \ln p(\mathbf{x}(t), \mathbf{x}(t-1), \mathbf{x}(t-2), ...; W, \mathbf{a}_1, ..., \mathbf{a}_{d_s})$. The resulting update equations are,

$$\Delta a_i(\tau) = -\mu u_i(t)\, s_i(t-\tau), \text{ with } u_i(t) = \frac{\partial \ln p(e_i(t))}{\partial e_i(t)} \tag{37}$$

The example in figure 4 serves as a demonstration of the method. Two independently uniformly distributed random variables where low-pass filtered and then linearly mixed to produce the distribution shown to the left. Additionally to the LPC coefficients each source coordinate was modeled by a Gaussian mixture density $p(e) = p(e|\alpha)$ with parameters α to allow for arbitrary source distributions. Overall the parameters estimated where therefore the linear mixture parameters W, the LPC coefficients for each source $\mathbf{a}_i$ and the parameters α_i of the Gaussian mixture model for each source. After estimating all this parameters with the outlined ML approach the original uniformly distributed coordinates and the proper statistical independent orientations have been recovered as seen in the center and to the right.

Figure 5 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. The density function for each channel was chosen as a zero mean Gaussian with unit variance. Gradient ascent rules (8) and (37) where used. The remaining cross-talk was hardly audible and corresponds to a signal-to-noise ratio (SNR) between 10 dB and 20 dB for the 10 different channels.

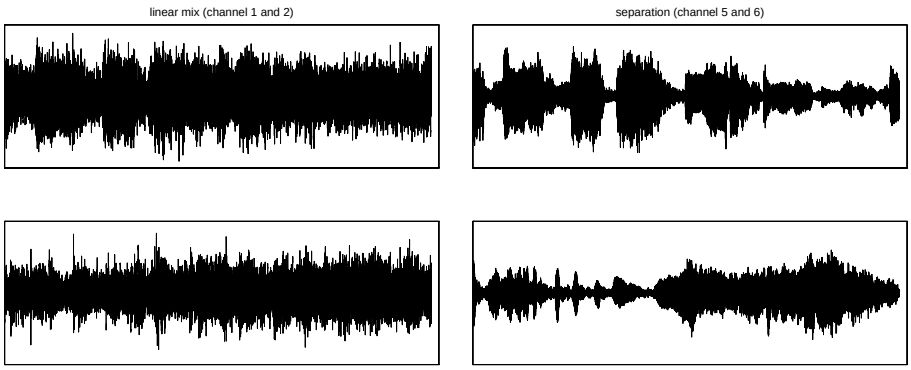

Figure 5: left: Two of the 10 channels of 10 linearly mixed music CD sources. right: Two channels of the output show good separation using contextual ICA. For more details see [21]

3.2 Separation based on time structure of cross-correlation

In this section we will see that using the temporal properties of the signal it is sufficient to consider second order statistic in oder to identify an unknown linear mixing. This is strictly speaking not ICA, which is defined as a factorization in all orders, but it can still be understood as source separation.

Again, for an instantaneous mixture the forward model is given by,

$$\mathbf{x}(t) = A\mathbf{s}(t) \tag{38}$$

We can formulate the cross-correlations $R_x(t,\tau)$ of the measured signals at time t and time delay τ

$$\begin{aligned} R_x(t,\tau) \equiv & \left\langle \mathbf{x}(t)\mathbf{x}^\top(t+\tau) \right\rangle \\ = & A \left\langle \mathbf{s}(t)\mathbf{s}^\top(t+\tau) \right\rangle A^\top \equiv A\Lambda_s(t,\tau)A^\top \end{aligned} \tag{39}$$

Since we assume decorrelated sources at all times we postulate diagonal cross-correlations $\Lambda_s(t,\tau)$. The notation is a bit vague here and needs some clarification. $R_x(t,0)$ represents the covariance estimated at time t, presumably estimated by averaging over a range around t. If we write $R_x(\tau)$ we are referring to the cross-correlation, or covariance with time delay τ estimated over all data, i.e. integrating over all t.

We now can choose to concentrate either on non-stationarity or non-whiteness of the signals.

In case of *non-stationary* signals we will exploit the fact that the covariance matrix estimated at different times t will give effectively new conditions for every time,

$$R_x(t,0) = \left\langle \mathbf{x}(t)\mathbf{x}^\top(t) \right\rangle = A\Lambda_s(t,0)A^\top \tag{40}$$

In the case of *non-white* signals we can exploit the fact that the cross-correlation (estimated over the entire signal) gives different conditions for different time delays τ

$$R_x(\tau) = \left\langle \mathbf{x}(t)\mathbf{x}^\top(t+\tau) \right\rangle = A\Lambda_s(\tau)A^\top \tag{41}$$

The diagonal terms in $\Lambda_s(\tau)$ represent now the auto-correlations of the model sources. But effectively, in both cases, by considering multiple times, or multiple delays respectively, we obtain a set of equations that have to be satisfied simultaneously. We have therefore a problem of simultaneous diagonalization.

For non-stationary signals a set of K equations (39) for different times $t_1, ...t_K$ gives then a total of $Kd_x(d_x+1)/2 + d_s$ constraint on $d_s d_x + d_s K$ unknown parameters $A, \Lambda_s(t_1), ..., \Lambda_s(t_K)$. [6]. Assuming all conditions are linearly independent[7] we will have sufficient conditions if,

$$Kd_x(d_x+1)/2 + d_s \geq d_s d_x + d_s K \tag{42}$$

In the square case, $d_s = d_x$, in principle $K = 2$ is sufficient to specify the solution up to arbitrary permutations. In that case the problem can be solved as a non-symmetric eigenvalue problem as outlined in [14].

$$\begin{aligned} R_x(t_1) &= A\Lambda(t_1)A^\top & (43)\\ R_x(t_2)^{-1} &= A^{-T}\Lambda(t_2)^{-1}A^{-1} & (44)\\ R_x(t_1)R_x(t_2)^{-1}A &= A\Lambda(t_1)\Lambda(t_2)^{-1} & (45) \end{aligned}$$

The last equation represents a non-symmetric eigenvalue problem. In general its solutions, A, are not orthogonal as expected. Though extremely fast the difficulty with such algebraic solutions, however, is that one does not have perfect estimates of $R_x(t)$, and the results depends strongly on the estimation noise. At best one can assume no-stationary signals and measure the sample estimates $\hat{R}_s(t)$ within the stationarity time. If we interpret the inaccuracy of that estimation as measurement error

$$E(k) = \hat{R}_x(k) - A\Lambda_s(k)A^\top \tag{46}$$

[6] We will write in the remainder in brief $\Lambda_s(k) = \Lambda_s(t_k)$ and $\Lambda_s = \Lambda_s(t_1), ..., \Lambda_s(t_K)$ whenever possible. The same applies to and $R_x(t)$

[7] Conditions on R_x and Λ_s for linear independence are outlined in [14].

it is reasonable to estimate the unknown parameters by minimizing the total measurement error for a sufficiently large K,

$$\hat{A}, \hat{\Lambda}_s = \underset{A,\Lambda_s,A_{ii}=1}{\arg\min} \sum_{k=1}^{K} \|E(k)\|^2 \tag{47}$$

The matrix norm here is the sum of the absolute values of every coefficient. Note that $\|E\|^2 = \mathrm{Tr}(EE^H)$. This represents a least squares (LS) estimation. To find the extrema of the LS cost $E = \sum_{k=1}^{K} \|E(k)\|$ in (47) let us compute the gradients with respects to its parameters[8]

$$\frac{\partial E}{\partial A} = -4 \sum_{k=1}^{K} E(k) A \hat{\Lambda}_s(k) \tag{48}$$

$$\frac{\partial E}{\partial \hat{\Lambda}_s(k)} = -2 \,\mathrm{diag}\left(AE(k)A^\top\right) \tag{49}$$

We will have to use an iterative algorithm to find the extrema with respect to A and $\Lambda_s(k)$ using the gradients in (48) and (49).

In the case of a square and invertible mixing $\hat{A}$ the signal estimates are trivially computed to be $\hat{\mathbf{s}} = \hat{A}^{-1}\mathbf{x}$. In the non-square case for $d_s < d_x$ we can compute the LS estimate,

$$\hat{\mathbf{s}}_{LS}(t) = \underset{\mathbf{s}(t)}{\arg\min} \|\mathbf{x}(t) - \hat{A}\mathbf{s}(t)\| = (\hat{A}^\top \hat{A})^{-1} \hat{A}^\top \mathbf{x}(t) \tag{50}$$

In this section we described how one can treat the case of instantaneous mixtures by decorrelating the covariance matrices simultaneously at several times. This approach requires non-stationary sources. The problem can also be treated by decorrelating the cross-correlation at different time delays. This requires the signals to be non-white rather than non-stationary. This is the approach traditionally take in the literature [2, 14, 23, 24].

3.3 Non-linear time dependencies

We discussed in the previous section only linear relations in time. In order to model non-linear relations we will get back to the general formulation of statistical independence with minimal mutual information of section 1.5. This section reproduces the work in [8]. Consider the problem of finding *non-linear* relations in a time series $x(1), ..., x(t), ...$. Very blandly speaking we know that non-linear recursion relations can lead to chaotic time series. We will use minimum mutual information to discover the

[8] The diagonalization operator here zeros the off-diagonal elements, i.e. $\mathrm{diag}(M)_{ij} = \begin{cases} M_{ij}, & i = j \\ 0, & i \neq j \end{cases}$

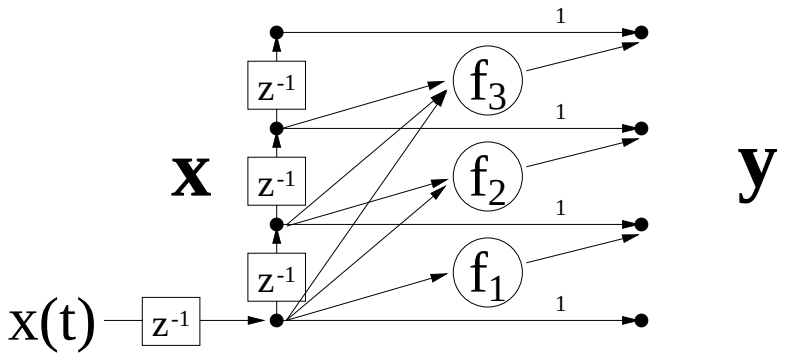

Figure 6: Every output of this network is connected to the past inputs with arbitrary non-linear relations parametrized by $\mathbf{w}$. Note that the direct connection weights are unity.

recursive relation leading to a chaotic time series. Consider the network structure of figure 6. Note that the output units are only connected to past inputs, and their connection to the present input is unit weight,

$$y(t) = x(t) - f(x(t-1), x(t-2), ...; \mathbf{w}) \tag{51}$$

With this we see that Jacobian has triangular structure, and the determinant is,

$$\left|\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right| = \begin{vmatrix} 1 & 0 & ... & 0 \\ ... & 1 & ... & 0 \\ ... & & ... & ... \\ ... & \partial \mathbf{f}(\mathbf{x};\mathbf{w})/\partial \mathbf{x} & 1 & 0 \\ ... & ... & & 1 \end{vmatrix} = 1 \tag{52}$$

The relation $f(\mathbf{x};\mathbf{w})$ can be any parametric non-linear transformation that best explains the relations of the time series. Due to this structure the map conserves the entropy and we obtain statistical independence if we minimize the single coordinate entropies. Note that this structure could be understood as a non-linear generalization of the traditional linear AR model. Equation (51) has the same structure as the linear AR equation (35), now however with non-linear relations to the past. The outputs $\mathbf{y}$ correspond to the error signal. Recall that the optimal linear AR parameter also lead decorrelated error signals, which represents statistically independent up to second order. We shall not go to far in this analogy and turn now to a demonstrative example. Consider for example

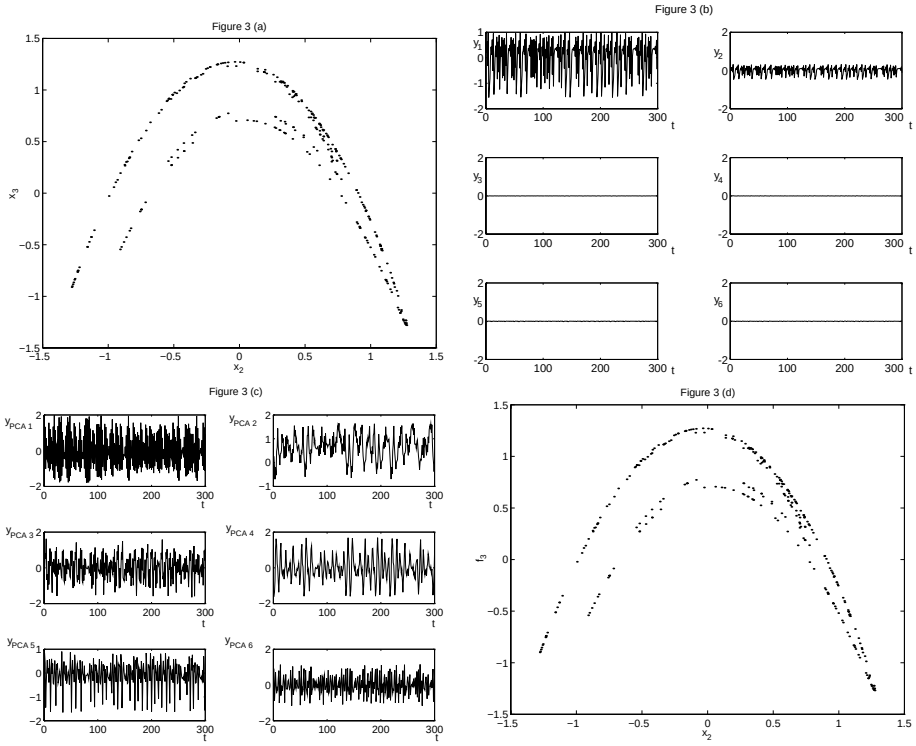

Figure 7: (a) State space structure of the attractor of the Henon map shown here by plotting $x(t+2)$ and $x(t+3)$. (b) outputs $y(t+1), y(t+2), ..., y(t+6)$ as a function of time after minimizing their mutual information, which reduces the correlation by extracting the functional relationship of the inputs $x(t+1), ..., x(t+6)$ (c) same as (b) but using standard PCA. (d) trained polynomial model f_3 versus $x(t+2)$. Figures taken from [7]

the Henon-map [11], which can be generated with the following iteration[9],

$$x_1(t) = 1 - 1.4x_1^2(t-1) + 0.3x_1(t-2) \tag{53}$$

To model this relation the output $y(t+i) = x(t+i) - f_i(x(t+i-1), ..., x(t+1), i = 1, ..., 6$, that is the prediction error, should be a polynomial combination of the past. By construction, however, the first output $y(t+1)$ is just $x(t+1)$. The polynomial of the following outputs cancel the effect of the iteration such that they are independent from the former. Minimizing the mutual information of the outputs will generate statistically independent output only if the model parameter match the parameters that generated the time series. In this example higher order statistics using cross-cumulants have been used to minimize mutual information. Figure 7 shows that the parameter of the polynomial f_3 match after training the parameters of (53). It also shows that a history of at least two taps is required to cancel the prediction errors in agreement with the generating dynamic. The polynomials f_4, f_5, f_6 have the same parameters than f_3 while zeroing all other parameters with longer history. They reveal thus the rank, or embedding dimension of the non-linear dynamic.

4 Convolutive BSS

In a real environment, where the signals travels slow compared to its correlation time, the instantaneous mix is not a good description of the linear superposition. The signal arrive at the different sensors with different time delays. In fact, the signal may be reflect at boundaries and arrive with multiple delays to a particular sensor. This scenario is referred to as a multi-path environment and can be described as a finite impulse response (FIR) convolutive mixture,

$$\mathbf{x}(t) = \sum_{\tau=0}^{P} A(\tau)\mathbf{s}(t-\tau) \tag{54}$$

How can one identify the $d_x d_s P$ coefficients of the channels A and how can one find an estimate $\hat{\mathbf{s}}(t)$ for the unknown sources? This situation is considerably more complicated than in the previous sections as one has now a matrix of filters rather than a matrix of scalars mixing. And even once the channel has been identified, inverting it is a more difficult task

[9] Note that this is not the original definition of the Henon-map, which is a nonlinear iteration of the sort, $\mathbf{z}(t+1)=\mathbf{f}(\mathbf{z}(t))$, with a two dimensional vector $\mathbf{z} \in \Re^2$. With the Takens's theorem, however, it is possible to convert any non-linear multi-dimensional dynamic $\mathbf{z}(t) \rightarrow \mathbf{z}(t+1)$ into a iteration of a one-dimensional dynamic with a longer history $x(t-P), ..., x(t) \rightarrow x(t+1)$. This, so called, embedding has the same dynamic properties, such as largest Lyapunov exponent and correlation dimension, than the original iteration [22]. The number of taps P is called the embedding dimension, and represents a important parameter for the modeling of chaotic time series.

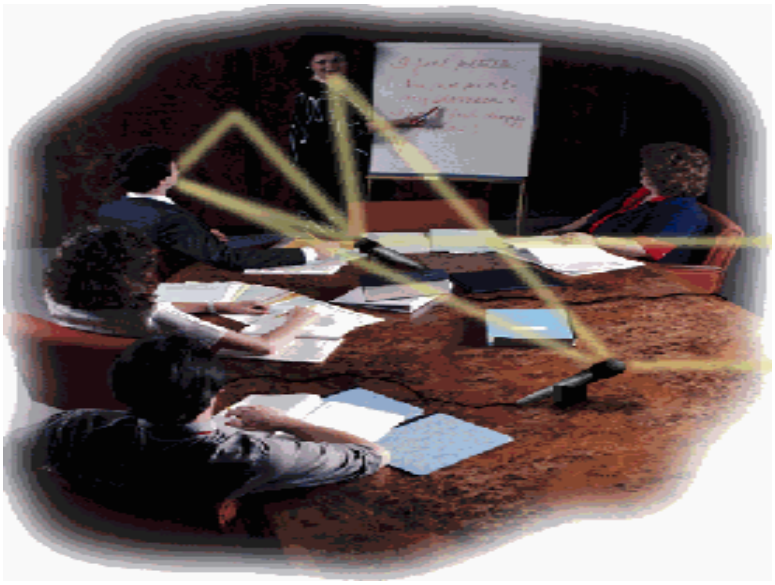

Figure 8: Acoustic signals recorded simultaneously in a reverberant environment can be described as sums of differently convolved sources.

as the inverse should in principle be a recursive, and therefore potentially an unstable infinite impulse response (IIR) filter.

Alternatively one may formulate an FIR inverse model W,

$$\mathbf{y}(t) = \sum_{\tau=0}^{Q} W(\tau)\mathbf{x}(t-\tau) \tag{55}$$

and try to estimate W such that the model sources $\mathbf{y}(t) = [y_1(t), ..., y_{d_y}(t)]^\top$ are statistically independent.

In analogy to section 3.2 we will exploit non-stationarity of the signals, which allows us to restrict to second order statistic. Higher order method for convolutive separation will not be discussed as they are substantially more complex and not really required for the case of non-stationary signals. The literature on convolutive BSS based on higher order methods is extensive. References can be found in [19].

4.1 Cross-correlations, circular and linear convolution

First consider the cross-correlations $R_x(t, t+\tau) = \left\langle \mathbf{x}(t)\mathbf{x}(t+\tau)^\top \right\rangle$. For stationary signals the absolute time does not matter and the correlations

depend on the relative time, i.e. $R_x(t, t+\tau) = R_x(\tau)$. Denote with $R_x(z)$ the z-transform of $R_x(\tau)$. We can then write

$$R_x(z) = A(z)\Lambda_s(z)A(z)^H \tag{56}$$

where $A(z)$ represents the matrix of z-transforms of the FIR filters $A(\tau)$, and $\Lambda_s(z)$ are the z-transform of the auto-correlation[10] of the sources, which again is diagonal due to the independence assumptions.

For practical purposes we have to restrict ourself to a limited number of sampling points of z. Naturally we will take T equidistant samples on the unit circle such that we can use the discrete Fourier transform (DFT). For periodic signals the DFT allows us to express circular convolutions as products such as in (56). However, in (54) and (55) we assumed linear convolutions. A linear convolution can be approximated by a circular convolution if $P \ll T$ and we can write approximately

$$\mathbf{x}(\omega, t) \approx A(\omega)\mathbf{s}(\omega, t), \text{ for } P \ll T \tag{57}$$

where $\mathbf{x}(\omega, t)$ represents the DFT of the frame of size T starting at t, $[\mathbf{x}(t), ..., \mathbf{x}(t+T)]$, and is given by $\mathbf{x}(\omega, t) = \sum_{\tau=0}^{T-1} e^{-\mathbf{i}2\pi\omega\tau}\mathbf{x}(t+\tau)$ and corresponding expressions for $\mathbf{s}(\omega, t)$ and $A(\omega)$.

For non-stationary signals the cross-correlation will be time dependent. Estimating the cross-correlation at the desired resolution of $1/T$ is difficult if the stationarity time of the signal is in the order of magnitude of T or smaller. We are content however with any cross-correlation average which diagonalizes for the source signals. One such sample average is,

---

[10] Recall the definition of the z transform

$$a(z) = \sum_{t=-\infty}^{\infty} a(t)z^{-t}$$

Read $a(z)$ to be the z transform of $a(t)$. This leads directly to the convolution theorem for the z-transform

$$y(t) = \sum_{\tau=-\infty}^{\infty} a(\tau)x(t-\tau) \iff y(z) = a(z)x(z)$$

since

$$\sum_{t=-\infty}^{\infty} z^{-t}y(t) = \sum_{t=-\infty}^{\infty} z^{-t} \sum_{\tau=-\infty}^{\infty} a(\tau)x(t-\tau) = \sum_{\tau=-\infty}^{\infty} z^{-\tau}a(\tau) \sum_{t'=-\infty}^{\infty} z^{-t'}x(t')$$

. Going back to the multi-path convolution (54) and inserting the definition for the cross-correlation matrix we find,

$$R_x(\tau) = \sum_{\tau'=-\infty}^{\infty} \sum_{\tau''=-\infty}^{\infty} A(\tau')\Lambda_s(\tau' - \tau'' - \tau)A^H(\tau'')$$

This gives with the convolution theorem equation (56)

$$\bar{R}_x(\omega,t) = \frac{1}{N}\sum_{n=0}^{N-1} \mathbf{x}(\omega,t+nT)\mathbf{x}^H(\omega,t+nT) \tag{58}$$

We can then write for such averages

$$\bar{R}_x(\omega,t) = A(\omega)\Lambda_s(\omega,t)A^H(\omega) \tag{59}$$

If N is sufficiently large we can assume that $\Lambda_s(\omega,t)$ can be modeled as diagonal again due to the independence assumption. For equations (59) to be linearly independent for different times t it will be necessary that $\Lambda_s(\omega,t)$ changes over time for a given frequency, i.e. the signal are non-stationary.

4.2 Backward model

Given a forward model A it is not guaranteed that we can find a stable inverse. In the two dimensional square case the inverse channel is easily determined from the forward model [24, 16]. It is however not apparent how to compute a stable inversion for arbitrary dimensions. In this present work we prefer to estimate directly a stable multi-path backward FIR model such as (55). In analogy to the discussion above and to section 4.2 we wish to find model sources with cross-power-spectra satisfying[11],

$$\Lambda_s(\omega,t) = W(\omega)\bar{R}_x(\omega,t)W^H(\omega) \tag{60}$$

In order to obtain independent conditions for every time we choose the times such that we have non-overlapping averaging times for $\bar{R}_x(\omega,t_k)$, i.e. $t_k = kTN$. But if the signals vary sufficiently fast overlapping averaging times could have been chosen. A multi-path channel W that satisfies these equations for K times simultaneously can be found, again with an LS estimation[12]

$$\begin{aligned} E(\omega,k) &= W(\omega)\bar{R}_x(\omega,k)W^H(\omega) - \Lambda_s(\omega,k) \\ \hat{W},\hat{\Lambda}_s,\hat{\Lambda}_n &= \underset{\substack{W,\Lambda_s \\ W(\tau)=0,\tau>Q, \\ W_{ii}(\omega)=1}}{\arg\min} \sum_{\omega=1}^{\top}\sum_{k=1}^{K} \|E(\omega,k)\|^2 \end{aligned} \tag{61}$$

[11] $W(\omega)$ represents the DFT with frame size T of the time domain $W(\tau)$. In what follows time and frequency domain are identified by their argument τ or ω.

[12] In short we write again $\Lambda_s(\omega,k) = \Lambda_s(\omega,t_k)$ and $\Lambda_s = \Lambda_s(\omega,t_1),...,\Lambda_s(\omega,t_K)$ whenever possible. The same applies to $R_x(\omega,t)$

Note the additional constraint on the filter size in the time domain. Up to that constraint it would seem the various frequencies $\omega = 1, ..., T$ represent independent problems. The solutions $W(\omega)$ however are restricted to those filters that have no time response beyond $\tau > Q \ll T$. Effectively we are parameterizing Td_sd_x filter coefficients in $W(\omega)$ with Qd_sd_x parameters $W(\tau)$. The LS solutions can again be found with a gradient descent algorithm. We will first compute the gradients with respect to the complex valued filter coefficients $W(\omega)$ and discuss their projections into the subspace of permissible solutions in the following section.

For any real valued function $f(\mathbf{z})$ of a complex valued variable $\mathbf{z}$ the gradients with respect to the real and imaginary part are obtained by taking derivatives formally with respect to the conjugate quantities $\mathbf{z}^*$ ignoring the non-conjugate occurrences of $\mathbf{z}^*$ [12].

$$\frac{\partial f(z)}{\partial \Re(z)} + i\frac{\partial f(z)}{\partial \Im(z)} = 2\frac{\partial f(z)}{\partial z^*} \tag{62}$$

Therefore the gradients of the LS cost in (61) are,

$$\frac{\partial E}{\partial W^*(\omega)} = 2\sum_{k=1}^{K} E(\omega,k)W(\omega)\bar{R}_x(\omega,k) \tag{63}$$

$$\frac{\partial E}{\partial \hat{\Lambda}_s^*(\omega,k)} = -\operatorname{diag}\left(E(\omega,k)\right) \tag{64}$$

With (64)=0 one can solve explicitly for parameters $\Lambda_s(\omega,k)$, while parameters $W(\omega)$ may be computed with a gradient descent rule.

4.3 Permutations and constraints

The above unconstrained gradients can not be used as such but have to be constrained to remain in the subspace of permissible solutions with $W(\tau) = 0$ for $\tau > Q \ll T$. This is important since it is a necessary condition for equations (60) to hold to a good approximation.

Additionally, not all possible permutations of frequencies will lead to FIR filters which satisfy that constrain. Note that any permutation of the coordinates for every frequency will lead to exactly the same error $E(\omega,k)$. The total cost will therefore not change if we choose a different permutation of the solutions for every frequency ω. Obviously those solutions will not all satisfy the condition on the length of the filter. Effectively, requiring zero coefficients for elements with $\tau > Q$ will restrict the solutions to be smooth in the frequency domain, e.g., if $Q/T = 8$ the resulting DFT corresponds to a convolved version of the coefficients with a *sinc* function 8 times wider than the sampling rate.

It is therefore crucial to enforce that constraint by starting the gradient algorithm with an initial point that satisfies the constraints, and then following the constrained gradient. The normalization condition that avoid

trivial solutions of the LS optimization have to be enforced simultaneously. The constrained gradients are obtained by applying the corresponding projection operators. The projection operator that zeros the appropriate delays for every channel $W_{ij} = [W_{ij}(0), ..., W_{ij}(\omega), ..., W_{ij}(T)]^\top$ is

$$P = FZF^{-1} \tag{65}$$

where the DFT is given by $F_{ij} = 1/\sqrt{T}e^{-\mathrm{i}2\pi ij}$, and Z is diagonal with $Z_{ii} = 1$ for $i < Q$ and $Z_{ii} = 0$ for $i \geq Q$. The projection operator that enforces unit gains on diagonal filters $W_{ii}(\omega) = 1$ is applied simply by setting the diagonal terms of the gradients to zero.

4.4 Performance metric

The main difficulty in assessing the quality of a separation from real recordings is that the true sources are generally not available.

We define as the Signal to Interference Ratio (SIR) of a signal $\mathbf{s}(t)$ in a multi-path channel $H(\omega)$ the total signal powers of the direct channel versus the signal power stemming from cross channels.

$$SIR[H, s] = \frac{\sum_\omega \sum_i |H_{ii}(\omega)|^2 \left\langle |s_i(\omega)|^2 \right\rangle}{\sum_\omega \sum_{i \neq j} \sum_j |H_{ij}(\omega)|^2 \left\langle |s_j(\omega)|^2 \right\rangle} \tag{66}$$

In the case of known channels and source signals we can compute the expressions directly by using a sample average over the available signal and multiplying the powers with the given direct and cross channel responses. In the case of unknown channel response and underlying signals we can estimate the direct powers (numerator) and cross-powers (denominator) by using alternating signals. We estimate the contributions of source j while source j is 'on' and all other sources are 'off'. During periods of silence, i.e. all sources are 'off' we can estimate background noise powers in all channels to subtract from the signal powers.

On real recordings we can measure the performance by hand-segmenting the signal of alternating speakers into speech and non-speech to obtain the 'on' and 'off' labels. With this algorithm one obtains a separation of 14 dB from a signal that originally had 0 dB SIR.

We have used artificial random filters in order to determine the dependency of the algorithm on the various parameters such as number of channels, number of sources, filter size, and required signal length. All experiments reported in figure 9 used mixtures with an SIR of roughly 0 dB as input. The direct forward channels where constant gain ($A_{ii}(\omega) = 1$) and the cross-channels where set in the time domain to zero mean, normal, random numbers. The deviation was adjusted to produce in average a SIR of 0dB. We used $K = 5$ in all cases.

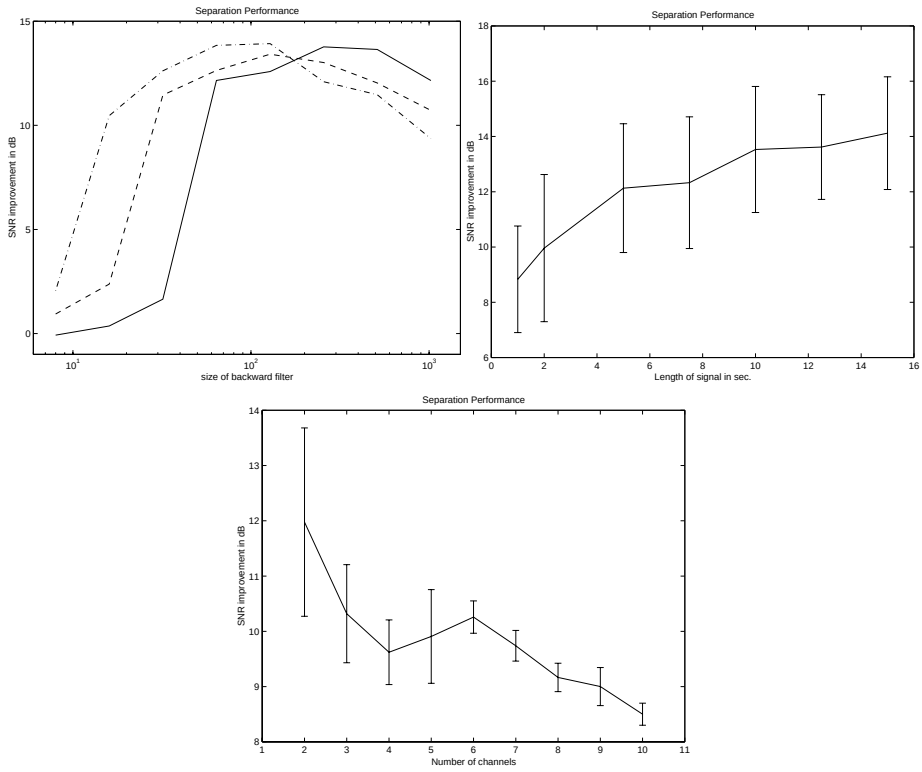

Figure 9: top left: Separation performance as a function of separation filter size Q for forward filter sizes $P = 16, 32, 64$ (dash-doted, dashed, and solid lines respectively). Mean values over 15 runs with different random forward filters are shown. The deviation from that mean was in average 3.4 dB, 4.8 dB, and 1.3 dB respectively. top right: Separation performance as a function of signal length in seconds for random forward filters of size $P = 64$. Mean and standard deviation over 15 runs with different random forward filters are shown. bottom: Separation performance as a function of number of channels ($d_s = d_x$) for random forward filters of size $P = 64$. Mean and standard deviation over 10 runs with different random forward filters are shown.

5 Summary

In a particular application the question arises which, if any, of these source separation techniques should be used. First, it is important to confirm that the model of a linear mixture of independent sources really applies, and if the mixture is instantaneous or convolutive. The cases of non-linear mixtures are not well understood yet, and are an area of active research.

If the samples are drawn independently one will have to resort to the criteria of statistical independence. Any particular knowledge about the density function of the underlying sources will be useful in restricting the class of possible non-linearities or equivalently the family of model distributions. If nothing, or very little is know about the underlying distributions higher moments may have to be considered.

If however the samples are given in a ordered sequence, and as a result are not drawn i.i.d., one can use the more robust second order techniques based on non-stationarity, or temporal correlation, the latter corresponding to spatial correlation in images.

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] Y. Bar-Ness, J. Carlin, and M. Steinberger. Bootstrapping adaptive cross-pol canceller for satellite communications. In *IEEE Int. Conf. Comunications*, pages 4F.5.1–4F.5.5, Philadelphia, PA, June 1982.

[3] A. Bell and T. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[4] P. Comon. Independent comonent analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

[5] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

[6] Gardiner C.W. *Handbook of Stochastic Methods. Second Edition.* Springer-Verlag, New York, 1990.

[7] G. Deco and Dragan Obradovic. *An Information Theoretic Approach to Neural Computing*. Perspective in Neural Computing. Springer, 1996.

[8] G. Deco and B. Schuermann. Learning time series evolution by unsupervised extraction of correlations. *Physical Review E*, 41:1780–1790, 1995.

[9] R. Duda and P. Hart. *Pattern classification and scene analysis*. John Wiley & Sons, 1973.

[10] H. Hayes, Monson. *Statistical Digital Signal Processing and Modeling*. Wiley, 1996.

[11] M. Henon. A two-dimensional mapping with a strange attractor. *Communications in Mathematical Physics*, 50(69), 1976.

[12] Klaus Jänich. *Einführung in die Funktionentheorie*. Springer-Verlag, 1977. see chaper 2 on Wirtinger-Kalkül.

[13] M.G. Kendal and A. Stuart. *The Advanced Theory of Statistics*. Charles Griffin & Company Limited, London, 1969.

[14] L. Molgedey and G. Schuster, H. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72(23):3634–3637, 1994.

[15] 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.

[16] Hoang-Lan Nguyen Thi and Christian Jutten. Blind source separation for convolutive mixtures. *Signal Processing*, 45:209–229, 1995.

[17] L. Parra. Symplectic nolinear component analysis. In *Advances in Neural Information Processing Systems 1995*, pages 437–443, Boston, MA, 1996. MIT Press.

[18] L. Parra, G. Deco, and S. Miesbach. Redundancy reduction with information preserving nonlinear maps. *Network: Computation in Neural Systems*, 6:61–72, 1995.

[19] Lucas Parra and Clay Spence. Convolutive blind source separation based on multiple decorrelation. *IEEE Speech and Audio Processing*, 1998. submitted.

[20] B. Pearlmutter and L. Parra. A context-sencitive generalization of independent component analysis. In *International Conf. on Neural Information Processing*, Hong Kong, 1996.

[21] B. Pearlmutter and L. Parra. Maximum likelihood source separation: a context-sensitive generalization of ica. In *NIPS 96*, Hong Kong, 1997.

[22] F. Takens. Detecting strange attractors in turbulence. In D.A. Rand and L.S. Young, editors, *Dynamical Systems and Turbulence*, volume 898 of *Lecture Notes in Mathematics*, pages 366–381. Spriger-Verlag, 1980.

[23] Stefan Van Gerven and Van Compernolle Dirk. Signal separation by symetric adaptive decorrelation: Stability, convergence, and uniqueness. *IEEE Transaction on Signal Processing*, 43(7):1602–1612, July 1995.

[24] 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.

[25] D. Yellin and E. Weinstein. Multichannel signal separation: Methods and analysis. *IEEE Transaction on Signal Processing*, 44(1):106–118, 1996.