The final version of this article has been published in
Neural Computation, 19(4):994-1021 (2007) published by The MIT Press.
This version does not differ significantly from the final version.

# Independent Slow Feature Analysis and Nonlinear Blind Source Separation

**Tobias Blaschke, Tiziano Zito and Laurenz Wiskott**
Institute for Theoretical Biology, Humboldt University Berlin
Invalidenstraße 43, D-10115 Berlin, Germany
{t.blaschke,t.zito,l.wiskott}@biologie.hu-berlin.de
http://itb.biologie.hu-berlin.de/~{blaschke,zito,wiskott}

**Abstract**

In the linear case statistical independence is a sufficient criterion for performing blind source separation. In the nonlinear case, however, it leaves an ambiguity in the solutions that has to be resolved by additional criteria. Here we argue that temporal slowness complements statistical independence well and that a combination of the two leads to unique solutions of the nonlinear blind source separation problem. The algorithm we present is a combination of second-order Independent Component Analysis and Slow Feature Analysis and is referred to as Independent Slow Feature Analysis. Its performance is demonstrated on nonlinearly mixed music data. We conclude that slowness is indeed a useful complement to statistical independence but that time-delayed second-order moments are only a weak measure of statistical independence.

## 1 Introduction

In signal processing one often has to deal with multivariate data such as a vectorial signal $\mathbf{x}(t) = [x_1(t), \ldots, x_M(t)]^\mathrm{T}$. To facilitate the interpretation of such a signal a useful representation of the data in terms of a linear or nonlinear transformation has to be found; prominent linear examples are Fourier transformation, Principal Component Analysis, and Fisher Discriminant Analysis. In this paper we will concentrate on Blind Source Separation (BSS), which recovers signal components (sources) that have originally generated an observed mixture. While the linear BSS problem can be solved by resorting to Independent Component Analysis (ICA), a method based on the assumption of mutual independence between the mixed source signal components, this is not possible in the nonlinear case. Some algorithms have been proposed to address this problem, and we will shortly mention them below. The objective of this paper is to show that the nonlinear BSS problem can be solved by combining ICA and Slow Feature Analysis (SFA), a method to find a representation where signal components are varying slowly.

After a short introduction to linear BSS and ICA in Section 2.1, we present the nonlinear BSS problem and some of the available algorithms in Section 2.2. SFA is explained in Section 3. We introduce Independent Slow Feature Analysis (ISFA) in Section 4, a combination of second-order ICA and SFA that can perform nonlinear BSS. In Section 5 the algorithm is tested on random and surrogate correlation matrices and then applied to nonlinearly mixed audio data. An analysis of the results reveals that nonlinear BSS can be solved by combining the objectives statistical independence and slowness, but that time-delayed second-order moments are not a sufficient measure of statistical independence in our case. We conclude with a discussion in Section 6.

# 2 Blind source separation and Independent Component Analysis

## 2.1 Linear BSS and ICA

Let $\mathbf{x}(t) = [x_1(t), \ldots, x_N(t)]^{\mathrm{T}}$ be a linear mixture of a source signal $\mathbf{s}(t) = [s_1(t), \ldots, s_N(t)]^{\mathrm{T}}$ and be defined by

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

with an invertible $N \times N$ mixing matrix $\mathbf{A}$. The goal of Blind Source Separation (BSS) is to recover the unknown source signal $\mathbf{s}(t)$ from the observable $\mathbf{x}(t)$ without any prior information. The only assumption is that the source signal components are statistically independent. Given only the observed signal $\mathbf{x}(t)$ we want to find a matrix $\mathbf{R}$ such that the components of

$$\mathbf{u}(t) = \mathbf{Q}\mathbf{y}(t) = \mathbf{Q}\mathbf{W}\mathbf{x}(t) = \mathbf{R}\mathbf{x}(t), \tag{2}$$

are mutually statistically independent. Here we have divided $\mathbf{R}$ into two parts. First a whitening transformation $\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t)$ with whitening matrix $\mathbf{W}$ is applied, resulting in uncorrelated signal components $y_i(t)$ with unit variance and zero mean, where we have assumed $\mathbf{x}(t)$ and also $\mathbf{s}(t)$ to have zero mean. In a second step a transformation $\mathbf{u}(t) = \mathbf{Q}\mathbf{y}(t)$ with orthogonal $\mathbf{Q}$ [Comon, 1994] results in statistically independent components $u_i(t)$.

The method of finding a representation of the observed data such that the components are mutually statistically independent is called Independent Component Analysis (ICA). It has been proven that ICA solves the linear BSS problem, apart from the fact that the source signal components can only be recovered up to scaling and permutation [Comon, 1994].

There exists a variety of algorithms performing linear ICA and therefore linear BSS. They can be divided into two classes [Cardoso, 2001]: (i) independence is achieved by optimizing a criterion that requires higher order statistics; (ii) the optimization criterion requires auto-correlations or non-stationarity of the source signal components. For the second class of BSS algorithms second-order statistics is sufficient [see e.g. Tong et al., 1991].

Here we focus on class (ii) and use a method introduced by Molgedey and Schuster [1994] based only on second-order statistics. It is based on the minimization of an objective function that can be written as

$$\Psi^{\tau}_{\mathrm{ICA}}(\mathbf{Q}) := \sum_{\substack{i,j=1\\ i\neq j}}^{N} \left(C^{(\mathbf{u})}_{ij}(\tau)\right)^2 = \sum_{\substack{i,j=1\\ i\neq j}}^{N} \left(\sum_{k,l=1}^{N} Q_{ik}Q_{jl}C^{(\mathbf{y})}_{kl}(\tau)\right)^2 \tag{3}$$

operating on the already whitened signal $\mathbf{y}(t)$. $C^{(\mathbf{u})}_{ij}(\tau)$ is an entry of a symmetrized time delayed correlation matrix

$$\mathbf{C}^{(\mathbf{u})}(\tau) \quad := \quad \frac{1}{2}\langle \mathbf{u}(t)\mathbf{u}(t+\tau)^{\mathrm{T}} + \mathbf{u}(t+\tau)\mathbf{u}(t)^{\mathrm{T}}\rangle, \tag{4}$$

$$C^{(\mathbf{u})}_{ij}(\tau) \quad := \quad \frac{1}{2}\langle u_i(t)u_j(t+\tau) + u_i(t+\tau)u_j(t)\rangle, \tag{5}$$

and $\mathbf{C}^{(\mathbf{y})}(\tau)$ is defined correspondingly. Minimization of $\Psi^{\tau}_{\mathrm{ICA}}$ can be understood intuitively as finding an orthogonal matrix $\mathbf{Q}$ that diagonalizes the correlation matrix with time delay $\tau$. Since, because of the whitening, the instantaneous correlation matrix, which is simply the covariance matrix, is already diagonal, this results in signal components that are decorrelated instantaneously and at a given time delay $\tau$. This can be sufficient to achieve statistical independence [Tong et al., 1991]. Extending this method to several time delays is straightforward and provides greater robustness, see e.g. [Belouchrani et al., 1997; Ziehe and Müller, 1998] and Section 5.1.

## 2.2 Nonlinear BSS and ICA

An obvious extension to the linear mixing model (1) has the form

$$\mathbf{x}(t) = F(\mathbf{s}(t)), \tag{6}$$

with a nonlinear function $F : \mathbb{R}^N \to \mathbb{R}^M$ that maps $N$-dimensional source vectors $\mathbf{s}(t)$ onto $M$-dimensional signal vectors $\mathbf{x}(t)$. The components $x_i(t)$ of the observable are a nonlinear mixture of the sources and like in the linear case source signal components $s_i(t)$ are assumed to be mutually statistically independent. Extracting the source signal is only possible if $F$ is an invertible function on the range of $\mathbf{s}(t)$, which we will assume from now on.

The equivalence of BSS and ICA in the linear case does not hold in general for a nonlinear function $F$ [Hyvärinen and Pajunen, 1999; Jutten and Karhunen, 2003]. For example, given statistically independent components $u_1(t)$ and

$u_2(t)$, any nonlinear functions $h_1(u_1)$ and $h_2(u_2)$ also lead to components that are statistically independent. Also a nonlinear mixture of $u_1(t)$ and $u_2(t)$ can still have statistically independent components [Jutten and Karhunen, 2003]. Thus in the nonlinear BSS problem independence is not sufficient to recover the original source signal and additional assumptions about the mapping $F$ or the source signal are needed to sufficiently constrain the optimization problem. We list some of the known methods:

- Constraints on the mapping $F$:
  - $F$ is a smooth mapping [Hyvärinen and Pajunen, 1999; Almeida, 2004];
  - $F$ is a post nonlinear (PNL) mapping [Taleb and Jutten, 1997; Yang et al., 1998; Taleb and Jutten, 1999; Taleb, 2002; Ziehe et al., 2003].
- Prior information about the source signal components:
  - source signal components are bounded [Babaie-Zadeh et al., 2002];
  - source signal components have time-delayed auto-correlations (referred to as temporal correlations) [Hosseini and Jutten, 2003];
  - source signal components are those that exhibit a characteristic time structure (power spectra are pairwise different) [Harmeling et al., 2003].

### 2.3 A new approach

In our approach we do not make any specific assumption about the mapping $F$, although the function space available for unmixing will be finite-dimensional in the algorithm, which imposes some limitations on $F$. Since we employ an ICA method based on time-delayed cross-correlations we make the implicit assumption that the sources have significantly different temporal structure (power spectra are pairwise different) [cf. Harmeling et al., 2003]. We also assume that the sampling rate is high enough, so that the input signal can be treated as if it were continuous and the time derivative is well approximated by the difference of two successive time points.

We have seen above, that in the nonlinear case statistical independence alone is not a sufficient criterion for blind source separation. There are infinitely many nonlinearly distorted versions of one source that are all statistically independent of another source. We propose slowness as a means to resolve this ambiguity and select a good representative from all the different versions of a source, because nonlinearly distorted versions of a source are usually varying more quickly than the source itself. Consider for example a sinusoidal signal component $x_i(t) = \sin(t)$ and a second component that is the square of the first $x_j(t) = x_i(t)^2 = 0.5\,(1-\cos(2t))$. The second component is more quickly varying due to the frequency doubling induced by the squaring. We believe this argument can be made more formal and it can be proven that, given the set of a one-dimensional signal and all its nonlinearly and continuously transformed versions, the slowest signal of the set is either the signal itself or an invertibly transformed version of it [Zito and Wiskott, in preparation]. Considering this we propose, in order to perform nonlinear BSS, to complement the independence objective of pure ICA with a slowness objective. In the next section we will give a short introduction to Slow Feature Analysis, an algorithm built on the basis of this slowness objective.

## 3 Slow Feature Analysis

Slow Feature Analysis (SFA) is a method that extracts slowly varying signals from a given observed signal [Wiskott and Sejnowski, 2002]. This section gives a short description of the method as well as a link between SFA and second-order ICA [Blaschke et al., 2006], which provides the means to find a simple objective function for our nonlinear BSS method.

Consider a vectorial input signal $\mathbf{x}(t) = [x_1(t), \ldots, x_M(t)]^{\mathrm{T}}$. The objective of SFA is to find a nonlinear input-output function $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), \ldots, g_L(\mathbf{x})]^{\mathrm{T}}$ such that the components of $\mathbf{u}(t) = \mathbf{g}(\mathbf{x}(t))$ are varying as slowly as possible. As a measure of slowness we use the variance of the first derivative, so that a slow signal has on average a small slope. The optimization problem then is as follows: Minimize the objective function

$$\Delta(u_i) \quad := \quad \langle \dot{u}_i^2(t) \rangle \tag{7}$$

successively for each $u_i(t)$ under the constraints

$$\begin{aligned}
\langle u_i(t)\rangle &= 0 && \text{(zero mean)}, && (8)\\
\langle (u_i(t))^2\rangle &= 1 && \text{(unit variance)}, && (9)\\
\langle u_i(t)u_j(t)\rangle &= 0 \ \forall\, j<i && \text{(decorrelation and order)}, && (10)
\end{aligned}$$

where $\langle\cdot\rangle$ denotes averaging over time. Constraints (8) and (9) ensure that the solution will not be the trivial solution $u_i(t) = \text{const.}$ Constraint (10) provides uncorrelated output signal components and thus guarantees that different components carry different information.

To make the optimization problem easier to solve we consider the components $g_i$ of the input-output function to be a linear combination of a finite set of nonlinear functions. We can then split the optimization procedure into two parts: (i) nonlinear expansion of the input signal $\mathbf{x}(t)$ into a high-dimensional feature space, and (ii) solving the optimization problem in this feature space linearly.

## 3.1 Nonlinear expansion

A common method to make nonlinear problems solvable in a linear fashion is nonlinear expansion. The observed signal components $x_i(t)$ are mapped into a high-dimensional feature-space according to

$$\mathbf{z}(t) = \mathbf{h}(\mathbf{x}(t)). \tag{11}$$

The dimension $L$ of $\mathbf{z}(t)$ is typically much larger than that of the original signal. For instance, if we want to expand into the space of second degree polynomials, we can apply the mapping

$$\mathbf{h}(\mathbf{x}) = [x_1, \ldots, x_M, x_1x_1, x_1x_2, \ldots, x_Mx_M]^{\mathrm{T}} - \mathbf{h}_0^{\mathrm{T}}. \tag{12}$$

The dimensionality of this feature space is $L = M + M(M+1)/2$. The constant vector $\mathbf{h}_0^{\mathrm{T}}$ is needed to make the expanded signal mean free.

## 3.2 Solution of the linear optimization problem

Given the nonlinear expansion, the nonlinear input-output function $\mathbf{g}(\mathbf{x})$ can be written as

$$\mathbf{g}(\mathbf{x}) \quad = \quad \mathbf{R}\mathbf{h}(\mathbf{x}) = \mathbf{R}\mathbf{z}\,, \tag{13}$$

where $\mathbf{R}$ is an $L \times L$ matrix which is subject to optimization. To simplify the optimization procedure we (i) choose the nonlinearities $\mathbf{h}(\cdot)$ such that $\mathbf{z}(t)$ is mean free and (ii) first find a transformation $\mathbf{y}(t) = \mathbf{W}\mathbf{z}(t)$ to obtain mutually decorrelated components $y_i(t)$ with zero mean. Matrix $\mathbf{W}$ is a whitening matrix as in normal ICA:

$$\mathbf{u}(t) = \mathbf{Q}\mathbf{y}(t) = \mathbf{Q}\mathbf{W}\mathbf{z}(t) = \mathbf{R}\mathbf{z}(t) = \mathbf{g}(\mathbf{x}(t))\,, \tag{14}$$

where $\mathbf{y}(t)$ is the nonlinearly expanded and whitened signal. It can be shown [Wiskott and Sejnowski, 2002] that the constraints (8), (9), and (10) are fulfilled trivially if the transformation $\mathbf{Q}$, subject to learning, is an orthogonal matrix. To solve the optimization problem we rewrite the slowness objective (7)

$$\Delta(u_i) = \langle(\dot{u}_1(t))^2\rangle = \mathbf{q}_i^{\mathrm{T}}\langle\dot{\mathbf{y}}(t)\,\dot{\mathbf{y}}(t)^{\mathrm{T}}\rangle\mathbf{q}_i =: \mathbf{q}_i^{\mathrm{T}}\mathbf{E}\mathbf{q}_i\,, \tag{15}$$

where $\mathbf{q}_i = [Q_{i1}, Q_{i2}, \ldots, Q_{iL}]^{\mathrm{T}}$ is the i-th row of $\mathbf{Q}$ and $\mathbf{E}$ is the matrix $\langle\dot{\mathbf{y}}(t)\dot{\mathbf{y}}(t)^{\mathrm{T}}\rangle$. For this optimization problem there exists a unique solution. For $i = 1$ the optimal weight vector is the normalized eigenvector that corresponds to the smallest eigenvalue of $\mathbf{E}$. The eigenvectors of the next higher eigenvalues produce the next slow components $u_2(t), u_3(t), \ldots$ and so forth. Typically only the first several of all $L$ possible output components are of interest and selected.

Finding the eigenvectors is equivalent to finding the transformation $\mathbf{Q}$ such that $\mathbf{Q}^{\mathrm{T}}\mathbf{E}\mathbf{Q}$ is diagonal. As described in detail in [Blaschke et al., 2006], this leads to an objective function for SFA subject to maximization

$$\Psi^{\tau}_{\mathrm{SFA}}(\mathbf{Q}) \quad := \quad \sum_{i=1}^{L}\left(C_{ii}^{(\mathbf{u})}(\tau)\right)^2 = \sum_{i=1}^{L}\left(\sum_{k,l=1}^{L} Q_{ik}Q_{il}C_{kl}^{(\mathbf{y})}(\tau)\right)^2\,, \tag{16}$$

where $\tau$ is a time delay that arises from an approximation of the time derivative. We set $\tau = 1$ because we make the approximation $\dot{\mathbf{y}}(t) \approx \mathbf{y}(t+1) - \mathbf{y}(t)$.

To understand (16) intuitively we note that slowly varying signal components are easier to predict and should therefore have strong correlations in time. Thus, maximizing the time delayed auto-correlation produces a slowly varying signal component. Since the trace of $\mathbf{C}^{(\mathbf{y})}(\tau)$ is preserved under a rotation $\mathbf{Q}$, maximizing the sum over the squared auto-correlations tends to produce a set of most slowly varying signal components at the expense of the other components, which become most quickly varying and are usually discarded.

Note the formal similarity between (3) and (16).

# 4 Independent Slow Feature Analysis

The nonlinear BSS method proposed in this section combines the principle of independence known from linear second-order BSS methods with the principle of slowness as described above. Because of the combination of ICA and SFA we refer to this method as Independent Slow Feature Analysis (ISFA). As already explained, second-order ICA tends to make the output components independent and SFA tends to make them slow. Since we are dealing with a nonlinear mixture we first compute a nonlinearly expanded signal $\mathbf{z}(t) = \mathbf{h}(\mathbf{x}(t))$ with $\mathbf{h} : \mathbb{R}^M \to \mathbb{R}^L$ being some nonlinear function chosen such that $\mathbf{z}(t)$ has zero mean. In a second step $\mathbf{z}(t)$ is whitened to obtain $\mathbf{y}(t) = \mathbf{W}\mathbf{z}(t)$. Finally we apply linear ICA combined with linear SFA on $\mathbf{y}(t)$ in order to find the output signal $\mathbf{u}(t)$, the $R$ first component of which are the estimated source signals, where $R$ is usually much smaller than $L$, the dimension of the expanded signal. Because of the whitening we know that ISFA, like ICA and SFA, is solved by finding an orthogonal $L \times L$ matrix $\mathbf{Q}$. We write the output signal $\mathbf{u}(t)$ as

$$\mathbf{u}(t) = \mathbf{Q}\mathbf{y}(t) = \mathbf{Q}\mathbf{W}\mathbf{z}(t) = \mathbf{Q}\mathbf{W}\mathbf{h}(\mathbf{x}(t)), \tag{17}$$

While the $u_1(t), \dots, u_R(t)$ are statistically independent and slowly varying, the components $u_{R+1}(t), \dots, u_L(t)$ are more quickly varying and may be statistically dependent on each other as well as on the estimated sources. The last $L-R$ components of the output signal $\mathbf{u}(t)$ are irrelevant for the final result but important during the optimization procedure, see below.

To summarize, we have an $M$-dimensional input $\mathbf{x}(t)$, an $L$-dimensional nonlinearly expanded and whitened $\mathbf{y}(t)$, and an $L$-dimensional output signal $\mathbf{u}(t)$. ISFA finds an orthogonal matrix $\mathbf{Q}$ such that the $R$ first components of the output signal $\mathbf{u}(t)$ are mutually independent and slowly varying. These are the estimated sources.

## 4.1 Objective function

To recover $R$ source signal components $u_i(t),\ i = 1, \dots, R$ from an $L$-dimensional expanded and whitened signal $\mathbf{y}(t)$ the objective for ISFA with one single time delay $\tau$ reads

$$\begin{aligned}\Psi^{\tau}_{\text{ISFA}}(u_1, \dots, u_R) &:= b_{\text{ICA}} \Psi^{\tau}_{\text{ICA}}(u_1, \dots, u_R) - b_{\text{SFA}} \Psi^{\tau}_{\text{SFA}}(u_1, \dots, u_R) \\ &= b_{\text{ICA}} \sum_{\substack{i,j=1, \\ i \neq j}}^{R} \left( C^{(\mathbf{u})}_{ij}(\tau) \right)^2 - b_{\text{SFA}} \sum_{i=1}^{R} \left( C^{(\mathbf{u})}_{ii}(\tau) \right)^2 ,\end{aligned} \tag{18}$$

where we simply combine the ICA objective (3) and SFA objective (16) for the first $R$ components weighted by the factors $b_{\text{ICA}}$ and $b_{\text{SFA}}$, respectively. Note that the ICA and the SFA objective are usually applied to all components and that in the linear case (and for one time delay $\tau = 1$) they are equivalent Blaschke et al. [2006]. Here, they are applied to an $R$-dimensional subspace in the $L$-dimensional expanded space, which makes them different from each other. $\Psi^{\tau}_{\text{ISFA}}$ is to be minimized, which is the reason why the SFA part has a negative sign.

In the linear case it is standard practice to use multiple time delays to stabilize the ICA solution, see for example the kTDSEP algorithm by Harmeling et al. [2003]. We will see in Sections 5.1 and 5.2 that in our case multiple time-delays are actually essential to get meaningful solutions. The general expression for the objective of ISFA then reads

$$\begin{aligned}\Psi_{\mathrm{ISFA}}(u_1,\ldots,u_R) &:= b_{\mathrm{ICA}}\sum_{\tau\in T_{\mathrm{ICA}}}\kappa^{\tau}_{\mathrm{ICA}}\Psi^{\tau}_{\mathrm{ICA}} - b_{\mathrm{SFA}}\sum_{\tau\in T_{\mathrm{SFA}}}\kappa^{\tau}_{\mathrm{SFA}}\Psi^{\tau}_{\mathrm{SFA}} \\ &= b_{\mathrm{ICA}}\sum_{\tau\in T_{\mathrm{ICA}}}\kappa^{\tau}_{\mathrm{ICA}}\sum_{\substack{i,j=1,\\ i\neq j}}^{R}\left(C^{(\mathbf{u})}_{ij}(\tau)\right)^2 \\ &\quad - b_{\mathrm{SFA}}\sum_{\tau\in T_{\mathrm{SFA}}}\kappa^{\tau}_{\mathrm{SFA}}\sum_{i=1}^{R}\left(C^{(\mathbf{u})}_{ii}(\tau)\right)^2 , \end{aligned} \tag{19}$$

where $T_{\mathrm{ICA}}$ and $T_{\mathrm{SFA}}$ are the sets of time delays for the ICA and SFA objectives respectively, whereas $\kappa^{\tau}_{\mathrm{ICA}}$ and $\kappa^{\tau}_{\mathrm{SFA}}$ are weighting factors for the corresponding correlation matrices. For simplicity we will first continue the description with only one time delay based on (18) and only later provide the full formulation with multiple time delays based on (19).

## 4.2 Optimization procedure

From (17) we know that $\mathbf{C}^{(\mathbf{u})}(\tau)$ in (18) depends on the orthogonal matrix $\mathbf{Q}$. There are several ways to find the orthogonal matrix that minimizes the objective function. Here we apply successive Givens rotations to obtain $\mathbf{Q}$. A Givens rotation is a rotation around the origin within the plane of two selected components $\mu$ and $\nu$ and has the matrix form

$$Q^{\mu\nu}_{ij} := \begin{cases} \cos(\phi) & \text{for } (i,j)\in\{(\mu,\mu),(\nu,\nu)\} \\ -\sin(\phi) & \text{for } (i,j)\in\{(\mu,\nu)\} \\ \sin(\phi) & \text{for } (i,j)\in\{(\nu,\mu)\} \\ \delta_{ij} & \text{otherwise} \end{cases} \tag{20}$$

with Kronecker symbol $\delta_{ij}$ and rotation angle $\phi$. Any orthogonal $L\times L$ matrix such as $\mathbf{Q}$ can be written as a product of $L(L-1)/2$ (or more) Givens rotation matrices $\mathbf{Q}^{\mu\nu}$ (for the rotation part) and a diagonal matrix with diagonal elements $\pm 1$ (for the reflection part). Since reflections do not matter in our case we only consider the Givens rotations as is often done in second-order ICA algorithms [e.g. Cardoso and Souloumiac, 1996] (but note that here it is applied to a subspace). The objective (18) as a function of a Givens rotation $\mathbf{Q}^{\mu\nu}$ reads

$$\begin{aligned}\Psi^{\tau,\mu\nu}_{\mathrm{ISFA}}(\mathbf{Q}^{\mu\nu}) &= b_{\mathrm{ICA}}\sum_{\substack{i,j=1\\ i\neq j}}^{R}\left(\sum_{k,l=1}^{L}Q^{\mu\nu}_{ik}Q^{\mu\nu}_{jl}C^{(\mathbf{u}')}_{kl}(\tau)\right)^2 \\ &\quad - b_{\mathrm{SFA}}\sum_{i=1}^{R}\left(\sum_{k,l=1}^{L}Q^{\mu\nu}_{ik}Q^{\mu\nu}_{il}C^{(\mathbf{u}')}_{kl}(\tau)\right)^2 , \end{aligned} \tag{21}$$

where $\mathbf{u}'$ is some intermediate signal during the optimization procedure. For each Givens rotation there exists an angle $\phi_{\min}$ with minimal $\Psi^{\tau,\mu\nu}_{\mathrm{ISFA}}$. Successive application of Givens rotations $\mathbf{Q}^{\mu\nu}$ with the corresponding rotation angle $\phi_{\min}$ leads to the final rotation matrix $\mathbf{Q}$ yielding

$$\mathbf{C}^{(\mathbf{u})}(\tau) = \mathbf{Q}^{\mathrm{T}}\mathbf{C}^{(\mathbf{y})}(\tau)\,\mathbf{Q}. \tag{22}$$

In the ideal case the upper left $R\times R$ submatrix of $\mathbf{C}^{(\mathbf{u})}(\tau)$ is diagonal with a large trace $\sum_{i=1}^{R} C^{(\mathbf{u})}_{ii}(\tau)$.

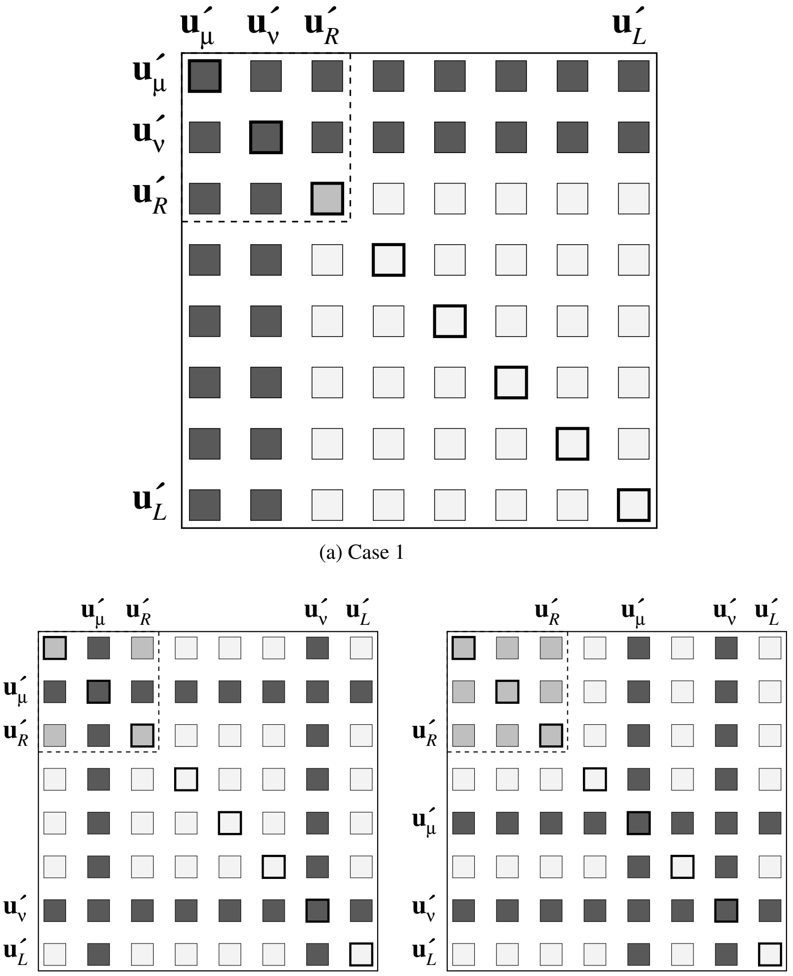
Applying a Givens rotation $\mathbf{Q}^{\mu\nu}$ in the $\mu\nu$-plane changes all auto- and cross-correlations $C^{(\mathbf{u}')}_{ij}(\tau)$ with at least one of the indices equal to $\mu$ or $\nu$. There exist two invariances under such a transformation, which can be described as

$$\left(C^{(\mathbf{u}')}_{\mu i}(\tau)\right)^2 + \left(C^{(\mathbf{u}')}_{\nu i}(\tau)\right)^2 = \text{const} \quad \forall \mathrm{i}\notin\{\mu,\nu\}, \tag{23}$$

$$\left(C^{(\mathbf{u}')}_{\mu\mu}(\tau)\right)^2 + \left(C^{(\mathbf{u}')}_{\mu\nu}(\tau)\right)^2 + \left(C^{(\mathbf{u}')}_{\nu\mu}(\tau)\right)^2 + \left(C^{(\mathbf{u}')}_{\nu\nu}(\tau)\right)^2 = \text{const}. \tag{24}$$

Assume we want to minimize $\Psi^{\tau}_{\mathrm{ISFA}}$ for a given $R$, where $R$ denotes the number of signal components we want to extract. Applying a Givens rotation $\mathbf{Q}^{\mu\nu}$ we have to distinguish three cases

- **Case 1** Both axes, $\mu$ and $\nu$, lie inside the subspace spanned by the first $R$ axes $(\mu,\nu \leq R)$ (see Fig. 1a): The sum over all squared cross correlations of all signal components that lie outside the $R$-dimensional subspace is constant as well as that of all signal components inside the subspace. The former holds because of the first invariance (23) and the latter because of the first (23) and second invariance (24). There is no interaction between inside and outside, in fact the objective function is exactly the objective for an ICA algorithm based on second-order statistics, e.g. TDSEP or SOBI [Ziehe and Müller, 1998; Belouchrani et al., 1997]. In [Blaschke et al., 2006] it has been shown that this is equivalent to SFA in the case of a single time delay of $\tau = 1$.



(a) Case 1

(b) Case 2

(c) Case 3

Figure 1: Each square represents a squared cross or auto-correlation $\left(C_{ij}^{(\mathbf{u}')}\right)^2$ where index $i\,(j)$ denotes the row (column) of the square. Dark squares indicate all entries that are changed by a rotation in the $\mu$-$\nu$-plane. $L$ is the dimensionality of the expanded signal $\mathbf{u}'$ and $R$ the number of signal components $u'_i(t)$ subject to optimization. The entries incorporated in the objective function are located in the upper left corner as indicated by the dashed line.

- **Case 2** Only one axis, w.l.o.g. $\mu$, lies inside the subspace; the other, $\nu$, lies outside $(\mu \leq R < \nu)$ (see Fig. 1b): Since one axis of the rotation plane lies outside the subspace, $u'_\mu$ in the objective function can be optimized at the expense of the $u'_\nu$ outside the subspace. A rotation of $\pi/2$, for example, would simply exchange components $u'_\mu$ and $u'_\nu$. For instance, according to (23) $\left(C_{\mu i}^{(\mathbf{u}')}\right)^2$ can be optimized at the expense of $\left(C_{\nu i}^{(\mathbf{u}')}\right)^2$ with $i \in \{1,\ldots,R\}$; according to (24) $\left(C_{\mu\mu}^{(\mathbf{u}')}\right)^2$ can be optimized at the expense of $\left(C_{\mu\nu}^{(\mathbf{u}')}\right)^2$, $\left(C_{\nu\mu}^{(\mathbf{u}')}\right)^2$, and $\left(C_{\nu\nu}^{(\mathbf{u}')}\right)^2$. This gives the possibility to find the slowest and most independent components in the whole space spanned by all $L$ axes in contrast to Case 1 where the minimum is searched within the subspace spanned by the first $R$ axes considered in the objective function.

- **Case 3** Both axes lie outside the subspace $(R < \mu, \nu)$ (see Fig. 1c):
  A Givens rotation with the two rotation axes outside the relevant subspace does not affect the objective function and can therefore be disregarded.

To optimize the objective function of ISFA (18) we need to calculate the explicit form of the objective function $\Psi^{\tau,\mu\nu}_{\mathrm{ISFA}}$ in (21). By inserting the Givens rotation matrix (20) into the objective function (21), and considering the case with multiple time delays, we can write the objective as a function of the rotation angle $\phi$

$$\begin{aligned}\Psi^{\mu\nu}_{\mathrm{ISFA}}(\phi) &= b_{\mathrm{ICA}}\left(e_c + \sum_{\beta=0}^{2} e_\beta \cos^{4-\beta}(\phi)\sin^{\beta}(\phi)\right) \\ &\quad - b_{\mathrm{SFA}}\left(d_c + \sum_{\alpha=0}^{4} d_\alpha \cos^{4-\alpha}(\phi)\sin^{\alpha}(\phi)\right) \qquad (25)\end{aligned}$$

with constants $e$ and $d$ that depend only on the $C^{(\mathbf{u}')}_{kl}$ before rotation. Further simplification [cf. Blaschke and Wiskott, 2004] leads to

$$\text{Case 1:} \quad \Psi^{\mu\nu}_{\mathrm{ISFA}}(\phi) = A_0 + A_4\cos(4\phi+\phi_4) \qquad (26)$$

$$\text{Case 2:} \quad \Psi^{\mu\nu}_{\mathrm{ISFA}}(\phi) = A_0 + A_2\cos(2\phi+\phi_2) + A_4\cos(4\phi+\phi_4) \qquad (27)$$

with a single minimum (if w.l.o.g. $\phi \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$), which can be calculated easily. The derivation of (26) and (27) involves various trigonometric identities and, because of its length, is documented in the appendix.

The iterative optimization procedure with successive Givens rotations can now be described as follows:

1. Initialize $\mathbf{Q}' = \mathbf{I}$ and compute $\mathbf{C}^{(\mathbf{u}')}(\tau) = \mathbf{C}^{(\mathbf{y})}(\tau)\,\forall\tau \in T_{\mathrm{ICA}} \cup T_{\mathrm{SFA}}$ with (4) and $\Psi'_{\mathrm{ISFA}}$ with (19).
2. Choose a random permutation of the set of axis pairs:
   $P = \sigma\left(\{(\mu,\nu), \text{ with } \mu \le R \text{ and } \mu < \nu \le L\}\right)$.
3. Go systematically through all axis pairs in $P$. For each axis pair:
   (a) determine the optimal rotation angle $\phi^{\mu\nu}_{\min}$ for the selected axes with (26) or (27),
   (b) compute the Givens rotation matrix $\mathbf{Q}^{\mu\nu}\left(\phi^{\mu\nu}_{\min}\right)$ defined by (20),
   (c) update $\mathbf{C}^{(\mathbf{u}')}(\tau)$ using $\mathbf{C}^{(\mathbf{u}')}(\tau) \to (\mathbf{Q}^{\mu\nu})^{\mathrm{T}}\mathbf{C}^{(\mathbf{u}')}(\tau)\mathbf{Q}^{\mu\nu}$,
   (d) update $\mathbf{Q}'$ according to $\mathbf{Q}' \to \mathbf{Q}^{\mu\nu}\mathbf{Q}'$,
   (e) backup the previous objective-function value $\Psi''_{\mathrm{ISFA}} = \Psi'_{\mathrm{ISFA}}$,
   (f) calculate the new objective-function value $\Psi'_{\mathrm{ISFA}}$ with (19) using the updated $\mathbf{C}^{(\mathbf{u}')}(\tau)$ from (3c),
   (g) store the relative decrease of the objective function value $\dfrac{\Psi''_{\mathrm{ISFA}} - \Psi'_{\mathrm{ISFA}}}{|\Psi''_{\mathrm{ISFA}}|}$.
4. Go to 2 until the relative decrease of the objective function is smaller than $\varepsilon \ll 1$ for all axis pairs in $P$.
5. Set $\mathbf{Q} = \mathbf{Q}'$ and $\mathbf{u}(t) = \mathbf{Q}\mathbf{y}(t)$.

In Step 2 it is important to note that the rotation planes of the Givens rotations are selected from the whole $L$-dimensional space (although we avoid the irrelevant Case 3 by requiring $\mu \le R$, see Fig. 1) whereas the objective function only uses information of correlations among the first $R$ signal components $u'_i$. Since $\mathbf{Q}_{\mu\nu}$ is very sparse, the Givens rotation in Step 3c does not require a full matrix-multiplication but can be computed more efficiently. Note that the algorithm works on the intermediate correlation matrices $\mathbf{C}^{(\mathbf{u}')}(\tau)$ and not on the signals themselves; the input signal $\mathbf{y}(t)$ is used only in the initialization (Step 1) and at the end (Step 5) when the output signal $\mathbf{u}(t)$ is computed. To circumvent the problem of getting stuck in local optima of the objective function, a random rotation of the outer space $(\nu > \mu > R)$ can be performed after convergence in Step 4, and the algorithm can be restarted at Step 2.

# 5 Results

To evaluate the performance of ISFA we tested the algorithm first on random matrices to check how many matrices are needed to get meaningful results, then on surrogate matrices to check that the algorithm reliably converges to the global optimum under these ideal conditions, and then on a difficult although low-dimensional mixture of audio data to show how it performs on real data. In order to reduce the problem of local optima, we use SFA as a preprocessing step. That choice follows from the empirical observation that SFA is always able to extract the first source signal. To stabilize the ISFA solutions even further we typically run the optimization routine once with the first axis fixed, and then once more following the procedure described in Section 4.2. Throughout the paper the SFA time-delay set and the weighting factors were as follows:

$$\begin{aligned} T_{\mathrm{SFA}} &= \{1\} && (28)\\ \kappa^{\tau}_{\mathrm{SFA}} &= 1 \quad \text{for } \tau = 1 && (29)\\ \kappa^{\tau}_{\mathrm{ICA}} &= 1 \quad \forall \tau \in T_{\mathrm{ICA}}\,; && (30) \end{aligned}$$

This particular choice makes it easy to interpret the ISFA objective function (19): The SFA part is the plain SFA objective function of (16); the ICA part is the plain ICA objective function of (3) extended to several time delays. If we would choose more than one time delay for the SFA part, the interpretation in terms of slowness would become less clear [see Blaschke et al., 2006]. $T_{\mathrm{ICA}}$ depends on the experiment, see below.

## 5.1 Tests with random matrices

First consider only the ICA part of the objective function (19). Its purpose is to guarantee statistical independence of the estimated sources by simultaneously diagonalizing the $R \times R$ upper left submatrix of $T$ time-delayed $L \times L$ correlation matrices, where $T$ is the number of elements in $T_{\mathrm{ICA}}$. However, for the ICA term to be useful we have to take sufficiently many matrices into account so that simultaneous submatrix-diagonalization is not trivial. For instance, a single symmetric matrix can always be fully diagonalized by the orthonormal set of its eigenvectors. Thus for $R = L$ and $T = 1$ one has to take at least two matrices to avoid this spurious solution, which would be found even if there are no underlying statistically independent sources.

To estimate the minimum number of matrices needed, we ran ISFA with $b_{\mathrm{SFA}} = 0$ on randomly generated symmetric matrices $\mathbf{A}^{\tau}$, $\tau = 1, \dots, T$, for different values of $L$, $R$, and $T$. The subdiagonalization was considered successful if $E := \sqrt{\langle A^2_{ij} \rangle_{\tau, j, i>j}}$, i.e. the square root of the averaged squared non-diagonal terms, was below a threshold $E_{\mathrm{crit}} := 10^{-3}$. For fixed $L$ and $R < L$ we typically observe that a high degree of subdiagonalization is possible for $T = 2$. For $T > 2$ the subdiagonalization is still possible but at a lower degree with increasing $T$, until a critical $T_{\mathrm{crit}}$ is reached, for which the degree of subdiagonalization displays a sharp transition where $E$ crosses the threshold $E_{\mathrm{crit}}$ and remains stable after that.

The estimated critical number of time delays $T_{\mathrm{crit}}$ for $L \in \{9, 20\}$ and different values of $R$ are given in Table 1. In the simulations that follow, we have $M = R = 2$ and use ISFA$^3$ and ISFA$^5$ (ISFA$^n$ refers to ISFA with polynomials of degree $n$) resulting in $L = 9$ and $L = 20$, respectively. From the table we see that with $T = 50$ we are well above $T_{\mathrm{crit}}$ in both cases.

| | $R$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | >10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $L = 9$ | $T_{\mathrm{crit}}$ | 18 | 8 | 5 | 4 | 3 | 2 | 2 | 2 | - | - |
| $L = 20$ | $T_{\mathrm{crit}}$ | 36 | 19 | 13 | 9 | 7 | 6 | 5 | 4 | 3 | 2 |

Table 1: Critical number of time delays, $T_{\mathrm{crit}}$, for different values of $L$ and $R$.

## 5.2 Tests with surrogate matrices

To test the performance of ISFA (now including the SFA part) in absence of noise, finite-size effects, or any other kind of perturbation we carried out an experiment with $T > 1$ surrogate matrices, prepared such that they have a unique exact solution (except for permutations). The first matrix, with $\tau = 1$, is fully diagonal with the diagonal elements ordered by decreasing absolute value, with the exception of the second and last element, which are swapped. All other $T - 1$ matrices are random symmetric matrices with a diagonal $(R + R_{\mathrm{ICA}}) \times (R + R_{\mathrm{ICA}})$ upper submatrix. SFA alone only sees the first matrix (cf. 28, 29) and would favor a solution in which the last component is swapped back into the $R \times R$ subspace in place of the small second component. ICA alone would favor any permutation of the first $(R + R_{\mathrm{ICA}})$ components equally well, because for any of these permutations the $R \times R$ upper submatrices are all diagonal. In this example ICA should prevent SFA from swapping the last component into the $R \times R$ subspace and SFA should disambiguate the many equally valid ICA solutions by selecting the largest diagonal elements, i.e. the slowest components, in the first matrix.

This set of matrices constitutes a fixed point for the ISFA algorithm. If we run ISFA directly on these matrices we get $\mathbf{Q} = \mathbf{I}$. If we now apply a random rotation matrix $\mathbf{Q}_{\mathrm{rand}}$ to the set of matrices, we would expect ISFA to find a matrix $\mathbf{Q}$ that inverts this rotation and returns the $R$ original first components, but in any arbitrary order. Thus, the $R \times R$ submatrix of the product $\mathbf{P} := \mathbf{Q}\mathbf{Q}_{\mathrm{rand}}$ should be a permutation matrix for perfect unmixing.

We performed 10,000 independent tests with $R = 2$, $R_{\mathrm{ICA}} = 2$, $L = 9$, and $T = 50$, somewhat imitating the case of two nonlinearly mixed independent sources and an expansion space of all polynomials of degree three. The estimated critical number of matrices $T_{\mathrm{crit}}$ is 18. Using 50 matrices we rule out any spurious solution as discussed in Section 5.1. As a measure of performance we used the reconstruction error measure first introduced by Amari et al. [1995] in the formulation given in [Blaschke and Wiskott, 2004]:

$$E = \frac{1}{R^2}\left(\sum_{i=1}^{R}\left(\sum_{j=1}^{R}\frac{|P_{ij}|}{\max_k |P_{ik}|} - 1\right) + \sum_{j=1}^{R}\left(\sum_{i=1}^{R}\frac{|P_{ij}|}{\max_k |P_{kj}|} - 1\right)\right). \tag{31}$$

An experiment is considered to be successful if the unmixing error is smaller than $10^{-5}$. We found that ISFA always recovered the original components and that this 100% success rate was largely independent of the scaling factors $b_{\mathrm{ICA}}$ and $b_{\mathrm{SFA}}$, which we therefore set to $b_{\mathrm{ICA}} = b_{\mathrm{SFA}} = 1$ for this experiment.

### 5.3 Tests with twisted audio data

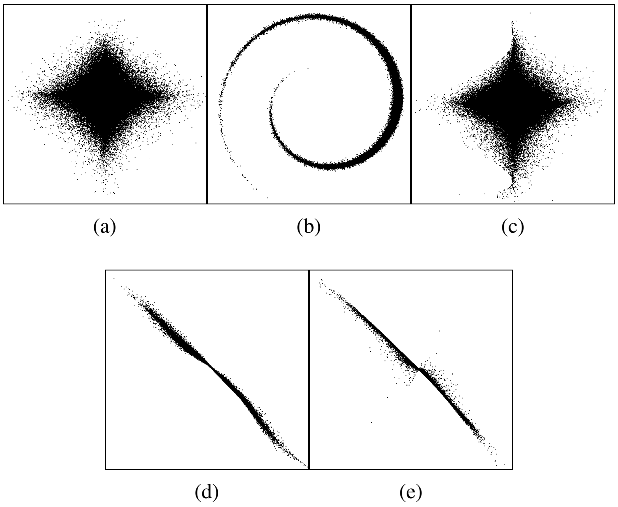
In the third experiment we tested the algorithm on 171 pairs of 19 nonlinearly mixed music excerpts. The sample values of the 19 excerpts were in the range of $[-1, +1)$; the mean had an average value of $(-10 \pm 110) \times 10^{-6}$ (mean ± std); the standard deviation had an average value of $0.16 \pm 0.07$, its minimum and maximum value was 0.02 and 0.27, respectively. One additional music excerpt was discarded, because it had extreme peaks, which led to a strong nonlinear distortion due to the SFA part and low correlations with the source even though it was in principle extracted correctly. All audio signals were $2^{21} = 2{,}097{,}152$ samples long and had a CD-quality sampling frequency of 44,100 Hz. We used the nonlinear mixture introduced by Harmeling et al. [2003] defined by

$$x_1(t) = (s_2(t) + 3s_1(t) + 6)\cos(1.5\pi s_1(t)), \tag{32}$$

$$x_2(t) = (s_2(t) + 3s_1(t) + 6)\sin(1.5\pi s_1(t)). \tag{33}$$

This is quite an extreme nonlinearity and the unmixing performance depends strongly on the standard deviation of the sources. For the ICA part of the objective in (19) we used 50 time delays evenly spaced within 1 and 44,100, corresponding to a time scale up to 1 second. The number of time delays is greater than the critical number $T_{\mathrm{crit}}$, which is 18 for an expansion with polynomials of degree three, and 36 for polynomials of degree five. In order to evaluate the performance of the algorithm fairly we used linear regression to check if the nonlinear mixture was indeed invertible within the available space. Two orthogonal directions were fit within the whitened expanded space to maximize the correlation with the original sources. Within the space of polynomials of degree three, there were a number of cases (51 examples, 30% of the total) where the two sources were not found by linear regression, which means the nonlinear mixture was not invertible within the available expanded space. This is the main reason for failures in ISFA[3]. Within the space of polynomials of degree five the mixture was always invertible. The scaling factor $b_{\mathrm{SFA}}$ was kept constant and equal to 1, while $b_{\mathrm{ICA}}$ was manually tuned for each example in order to maximize the correlation between estimated and original sources. For polynomials of degree three we tested different values of $b_{\mathrm{ICA}}$ equidistant on a logarithmic scale between 0 and 10000. The number of tested values varied between 5 and 40 depending on how clear and robust the optimum was. For polynomials of degree one and five we largely adopted the values found for polynomials of degree three; only if the algorithm failed with these values did we retune $b_{\mathrm{ICA}}$ with 20 equidistant values. This tuning resulted in values between 0 and 1000. A source signal is considered to be recovered if the correlation with the estimated source is greater than 0.9.

Scatter plots of a successful example are shown in Figure 2 and a summary of the results is given in Table 2. ISFA is able to separate the nonlinearly mixed sources in about 70% of the cases in which unmixing was possible at all. This is remarkable given the extreme nonlinearity of the mixture and a chance level of unmixing of less than 0.01%, as we have tested by numerical simulations. However, there remains a failure rate of about 30%, which is puzzling given the perfect performance on the surrogate matrices (Sec. 5.2). We investigate this in the next section.



(a) (b) (c)

(d) (e)

Figure 2: Scatter plot of two sources, their nonlinear mixture, and the estimated sources. (a) Sources, (b) mixture, (c) sources estimated by ISFA$^5$, (d) first source vs. estimated first source, (e) second source vs. estimated second source. Correlation coefficients of estimated sources and original sources were 0.996 and 0.998.

| # rec. src. | REG$^1$ | | REG$^3$ | | REG$^5$ | |
|---|---|---|---|---|---|---|
| 2 | 5% | (8) | 70% | (120) | 100% | (171) |
| 1 | 54% | (93) | 30% | (51) | 0% | (0) |
| 0 | 41% | (70) | 0% | (0) | 0% | (0) |
| # rec. src. | ISFA$^1$ | | ISFA$^3$ | | ISFA$^5$ | |
| 2 | 5% | (8) | 50% | (85) | 71% | (122) |
| 1 | 50% | (85) | 34% | (59) | 18% | (30) |
| 0 | 45% | (78) | 16% | (27) | 11% | (19) |
| **% correct** | **100%** | $\left(\frac{8}{8}\right)$ | **71%** | $\left(\frac{85}{120}\right)$ | **71%** | $\left(\frac{122}{171}\right)$ |

Table 2: The upper part shows percentages of cases where both, one, or none of the two sources were recovered by linear regression (supervised) in the original space (REG$^1$) or in the expanded space with polynomials of degree three (REG$^3$) or five (REG$^5$). The lower part shows the same for ISFA (unsupervised except for the tuning of $b_{\mathrm{ICA}}$). Each entry indicates the percentage (and number) of pairs with respect to the total of 171 pairs. The last line presents the percentage of both sources recovered correctly with respect to the number of mixtures invertible within the available expanded space by linear regression. Note that in the case of two recovered sources chance level is always smaller than 0.01%.

## 5.4 Analysis of failure cases

Why did ISFA fail in about 30% of the cases where a good solution was available by linear regression? The values of the objective function $\Psi_{\mathrm{ISFA}}$ (19) and its two parts $\Psi_{\mathrm{ICA}}$ and $\Psi_{\mathrm{SFA}}$ give us some information about possible reasons. Consider the following four different cases:

1. In 1 out of the 35 true failures for ISFA$^3$ and never for ISFA$^5$ it is the case that $\Psi_{\mathrm{ISFA}}$ of the sources estimated by ISFA is greater than the $\Psi_{\mathrm{ISFA}}$ of the sources estimated by linear regression. In this case the algorithm obviously got stuck in a local optimum.

2. In 15 and 26 out of the 35 and 49 true failures for ISFA$^3$ and ISFA$^5$, respectively, $\Psi_{\mathrm{ISFA}}$ of the sources estimated by ISFA is smaller than the $\Psi_{\mathrm{ISFA}}$ of the sources estimated by linear regression, but either $\Psi_{\mathrm{ICA}}$ or $\Psi_{\mathrm{SFA}}$ is greater than the corresponding linear-regression value. This indicates that the tuning of the weighting factors $b_{\mathrm{SFA}}$ and $b_{\mathrm{ICA}}$ might not have been fine enough. However, it could also be that there is an abrupt transition between solutions where $\Psi_{\mathrm{ICA}}$ is greater to solutions where $\Psi_{\mathrm{SFA}}$ is greater than the corresponding linear-regression value.

3. In 6 and 3 out of the 35 and 49 true failures for ISFA$^3$ and ISFA$^5$, respectively, $\Psi_{\mathrm{ICA}}$ and $\Psi_{\mathrm{SFA}}$ of the sources estimated by ISFA are both smaller than the ones of the linear-regression estimate and greater than the ones of the original sources. Neither a local optimum nor the weighting factors are a plausible cause for the failure in these cases. It might be that the expansion was too low-dimensional and that a higher-dimensional expansion would have yielded the correct solution.

4. In 13 and 20 out of the 35 and 49 true failures for ISFA$^3$ and ISFA$^5$, respectively, $\Psi_{\mathrm{ICA}}$ and $\Psi_{\mathrm{SFA}}$ of the sources estimated by ISFA are both smaller than the ones of the original sources. In this case the solution found is even better than the original sources in terms of the objective function, which indicates that there is something wrong with the objective function.

It might be possible to eliminate the failures of the first three cases by refining the algorithm, e.g. by tuning the weighting factors better or by going to higher polynomials, but Case 4 is more fundamental and requires to reconsider the objective function itself. In this latter case, the signals extracted by ISFA appear to be both slower *and* more mutually independent than the original sources. However, scatter plots of the estimated sources reveal that they are not statistically independent at all, but often one is largely a function of the other, see Figures 3 and 4. Thus the ICA part of the objective function is not strong enough to assure statistical independence of the estimated sources. The cross-correlation functions shown in Figure 5 indicate that this problem is not due to the specific choice of the time delays, because the time-delayed cross-correlations of the estimated sources (mean ± std = $0 \pm 0.0028$) are overall smaller than the ones of the original sources ($0 \pm 0.0066$). Even using different or more time delays, such dataset would have been processed incorrectly. We conclude that any measure of independence based on time-delayed correlations would be insufficient in our context.

Figure 5 suggested to us that sources with a large standard deviation of their cross-correlation function might be particularly difficult to separate with our ISFA algorithm. We tested this hypothesis but did not find a significant correlation with the failure cases. For an expansion with polynomials of degree three even linear regression fails if the standard deviation of the first signal, which goes along the spiral, is large. For polynomials of degree five linear regression always worked in our examples but we suspected that separation might still be more difficult for sources with large standard deviation, but again, we did not find a significant correlation with the failure cases.

We argue here that the failures must be attributed to the weakness of the ICA-term in the objective function. If the SFA-term were too weak, it could happen that all output signal components are truly statistically independent but at least some of them are too quickly varying, so that they are not correlated to the sources but to some nonlinearly distorted version of the sources, something we did not observe. Also the success in detecting the failure cases based on higher-order cumulants (see next section) indicates that the failures are due to the ICA-term.

## 5.5 Unsupervised detection of failure cases

A failure rate of about 30% (or even up to 50% for ISFA$^3$ if one also counts the cases in which even linear regression was not able to recover the sources) is obviously not acceptable, unless one can detect the failure cases in an unsupervised manner. We use the weighted sum over the third and fourth order cross-cumulants,

$$\Psi_{34}(\mathbf{u}) \quad := \quad \frac{1}{3!}\sum_{ijk\neq iii}\left(C^{(\mathbf{u})}_{ijk}\right)^2 + \frac{1}{4!}\sum_{ijkl\neq iiii}\left(C^{(\mathbf{u})}_{ijkl}\right)^2 , \tag{34}$$

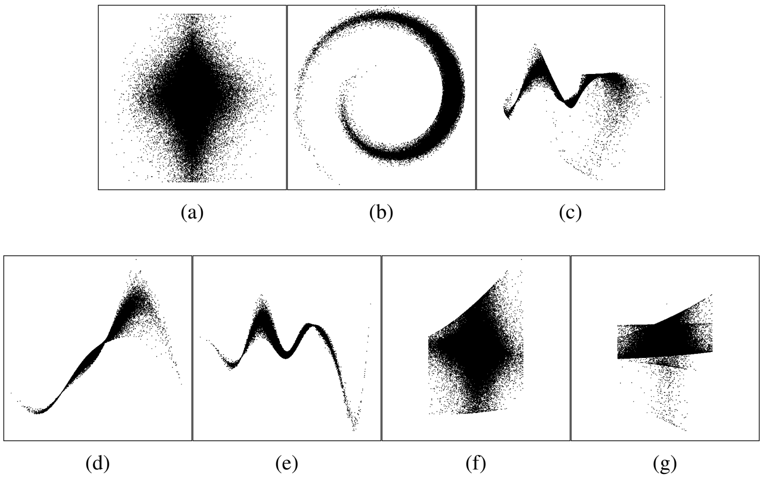


Figure 3: Scatter plot of two sources, their nonlinear mixture, and the sources estimated by ISFA in a failure case. (a) Sources, (b) mixture, (c) sources estimated by ISFA$^3$, (d) first source vs. estimated first source (corr. coeff. 0.9771), (e) first source vs. estimated second source (corr. coeff. 0.0377), (f) second source vs. estimated first source (corr. coeff. 0.0197), (g) second source vs. estimated second source (corr. coeff. 0.1301).

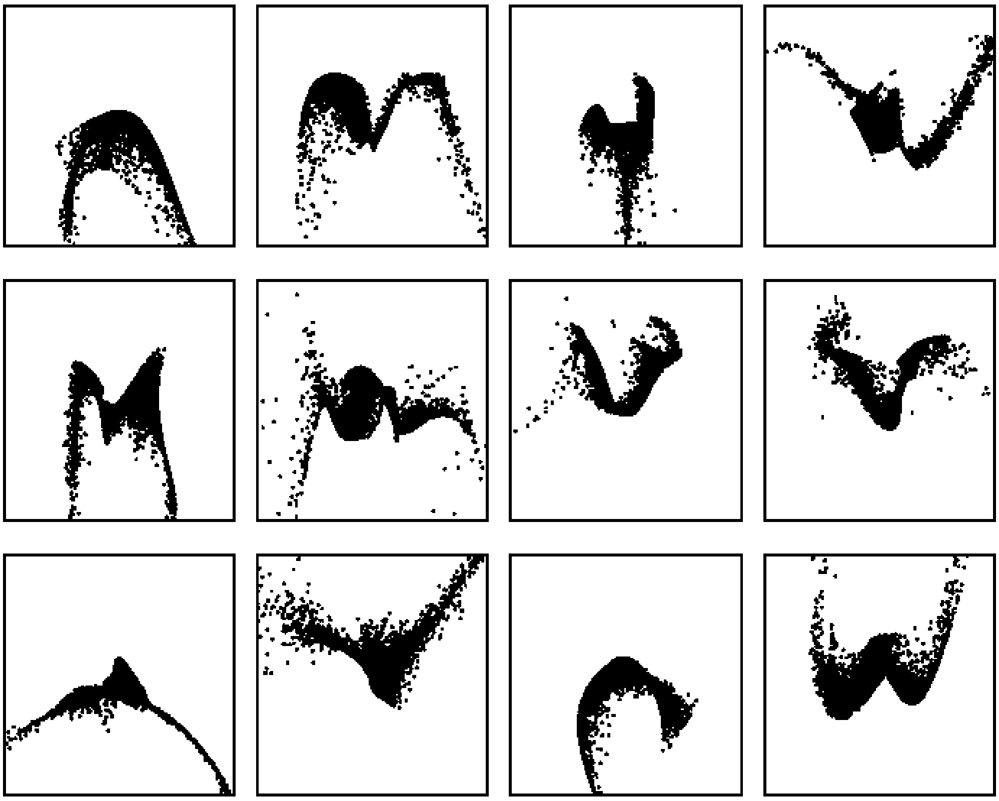
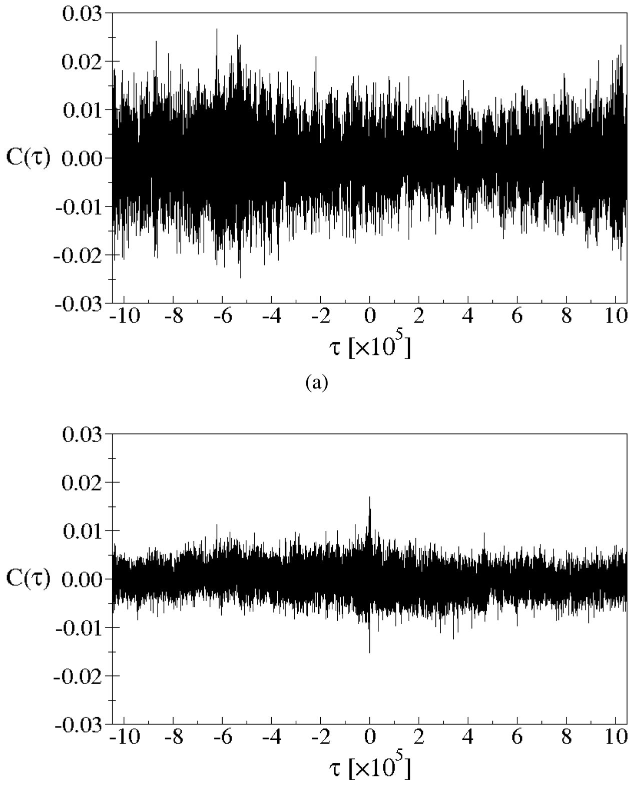


Figure 4: Scatter plots of the sources estimated by ISFA for some failure cases. It is clear that in these cases the signal components are not statistically independent even though the ICA-term indicates so.



(a)

(b)

Figure 5: Cross-correlation functions of a failure case: (a) cross-correlation function of the original sources, (b) cross-correlation function of the estimated sources. Same dataset as in Fig. 3.

as an independent measure of statistical independence to indicate with high values those cases in which the second order ICA-term has failed to yield independent output signal components. The factors $\frac{1}{3!}$ and $\frac{1}{4!}$ arise from an expansion of the Kullback-Leibler divergence in $\mathbf{u}$, which provides a rigorous derivation of this criterion [Comon, 1994; McCullagh, 1987]. The Receiver Operating Characteristic (ROC) curves in Figure 6 show that $\Psi_{34}(\mathbf{u})$ is a good measure of success. These tests also included the cases where linear regression was not able to recover the sources. The area under the ROC curves is 0.952 and 0.988 for ISFA$^3$ and ISFA$^5$, respectively.

# 6 Conclusion

In the work presented here we have addressed the problem of nonlinear blind source separation. It is known that in contrast to the linear case statistical independence alone is not a sufficient criterion for separating sources from a nonlinear mixture; additional criteria are needed to solve the problem of selecting the true sources (or good representatives thereof) from the many possible output signal components that would be statistically independent of other components. We claim here that for source signals with significant autocorrelations for time delay one temporal slowness is a good criterion to solve this selection problem, because the slow components are those most likely related to the true sources by an invertible transformation; non-invertible transformations would typically lead to more quickly varying components.

Based on this assumption, we have derived an objective function that combines a term from second-order Independent Component Analysis (ICA) with a term derived from Slow Feature Analysis (SFA). Optimization of the new objective function is achieved by successive Givens rotations, a method often used in context of ICA. We refer to the resulting algorithm as independent Slow Feature Analysis (ISFA) to indicate the combination of ICA and SFA.

The algorithm is somewhat unusual in that only a small submatrix of large time-delayed correlation matrices are being diagonalized by the Givens rotations (usually the full matrices are being diagonalized). This opens the question of the uniqueness of the solution. Using randomly generated pseudo-correlation-matrices we have found that indeed a minimum number of time delays is needed to obtain unique and meaningful solutions. For instance, if the upper left $2 \times 2$-submatrix of $9 \times 9$-matrices have to be diagonalized, at least 18 such matrices are needed to obtain a meaningful

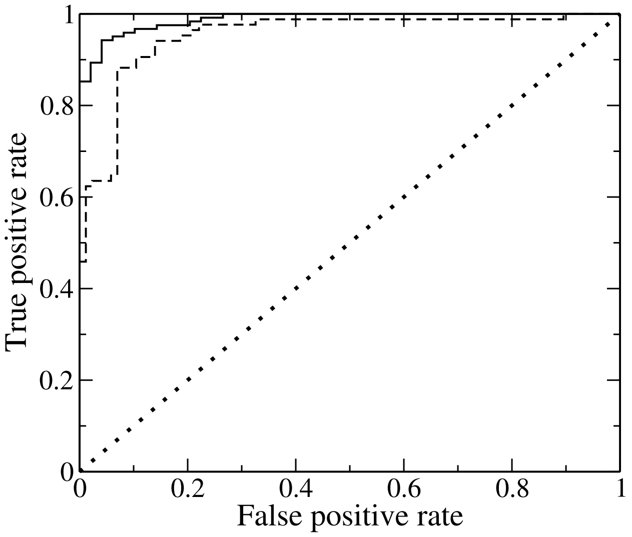


Figure 6: ROC curves for the test of successful source separation based on $\Psi_{34}(\mathbf{u})$, the weighted sum of third- and fourth-order cross-cumulants. The area under the curves is 0.952 and 0.988 for ISFA$^3$ (dashed line) and ISFA$^5$ (solid line), respectively.

solution that would be very unlikely to emerge by accident; with 17 matrices on the other hand good diagonalization can be achieved reliably even for random symmetric matrices. With (sufficiently many) surrogate matrices, structured such that they have a unique solution, we have subsequently verified that the algorithm reliably converges to the correct solution.

With tests on quite an extreme nonlinear mixture of two audio signals we have shown that ISFA is indeed able to perform nonlinear blind source separation, often with high precision. However, in about 30% of the cases in which the true sources could have been extracted with the nonlinearity used (as verified by regression) ISFA failed to extract them. In many of these cases the extracted signals were actually better than the original sources in both the SFA-term as well as the ICA-term of the objective function. This was a surprising finding for us, since it seems to contradict our basic assumption that a combination of slowness and statistical independence should permit reliable nonlinear blind source separation. Closer inspection, however, has revealed that the extracted output signal components only appear to be statistically independent in terms of the time-delayed second-order moments but that they are often highly related, as can be seen by visual inspection (Fig. 4) and automatically detected with a measure $\Psi_{34}(\mathbf{u})$ based on higher-order cumulants. This is not a consequence of the particular choice of time delays we have used but would be expected for any general set of time delays, as can be seen from the cross-correlation functions (Fig. 5).

We believe that two important conclusions can be drawn from these results. Firstly, the success cases indicate that combining slowness and statistical independence is a promising approach to nonlinear blind source separation. Secondly, any measure of statistical independence based on (time-delayed) second-order moments is too weak to guarantee statistical independence in our context; it might even be too weak in any context where the dimensionality of the space in which the signal components are searched for is significantly larger than the number of components.

For a possible theoretical account of the failure of second-order ICA in our context consider the following example. Given a symmetrically distributed source $s_1$ the correlation between, for instance, $s_1$ and $s_1^2$ vanishes [Harmeling et al., 2003, sec. 4.1]. To the extent that this also holds for time-shifted versions $s_1(t)$ and $s_1^2(t+\tau)$ [cf. Harmeling et al., 2003, sec. 5.4] the statistical dependence between $s_1$ and $s_1^2$ does not manifest itself in the time-delayed correlations. Thus, second-order ICA cannot be expected to prevent extraction of $s_1$ and $s_1^2$ as the estimated sources, which can easily lead to a failure case, if $s_1^2$ is more slowly varying than, e.g., $s_2$ .

A failure rate of 30% would render the algorithm useless if it were not possible to detect the failure cases. We have shown that the measure $\Psi_{34}(\mathbf{u})$, which is based on higher-order cumulants, permits failure detection with high reliability; the area under the ROC curve is greater than 0.95 resulting in a true positive rate of 90% and 94% at a false positive rate of 5% and 10% for ISFA$^3$ and ISFA$^5$, respectively.

It might be possible to use $\Psi_{34}(\mathbf{u})$ not only to detect the failure cases but also to automatically tune the weight $b_{\text{ICA}}$ given $b_{\text{SFA}} = 1$ and to determine the number of sources. For the former one could start with a small value of $b_{\text{ICA}}$, so that only the SFA-term is effective and the extracted components might not be independent, and then increase $b_{\text{ICA}}$, so that the ICA-term becomes increasingly effective, until the value of $\Psi_{34}(\mathbf{u})$ drops below a certain threshold. Similarly, for determining the number of sources one could start by running the algorithm with only two output components to be extracted and successively increase the number of components. One would then stop if adding another component would increase $\Psi_{34}(\mathbf{u})$ significantly (which can obviously be detected only *a posteriori*).

More interesting, however, would be to use higher-order cumulants more directly to improve the algorithm. For instance, one could define a new objective function that is a combination of the SFA-term used here and an ICA-term like $\Psi_{34}(\mathbf{u})$. Given the high reliability with which $\Psi_{34}(\mathbf{u})$ can detect failure cases, we expect better performance with such a new objective function. However, higher-order cumulants are expensive to compute, especially for high-dimensional and long signals, so that there is probably a trade-off between reliability and computational complexity. Exploring these possibilities will be subject of our future research.

## Acknowledgments

This work has been supported by the Volkswagen Foundation through a grant to LW for a junior research group. We thank Pietro Berkes for carefully going through all constants and Barak Pearlmutter for helpful comments. All examples have been implemented in Python using the Modular Toolkit for Data Processing[1].

## Appendix

The definitions of the constants $d_n$ and $e_n$ for the expression of the objective function (25) follow directly from the multilinearity of $C^{(\mathbf{u})}_{\cdots}(\tau)$. They are given in Table 3. Using trigonometry we can derive simpler objective functions of the form

$$\begin{aligned}\text{Case 1: } \Psi^{\mu\nu}_{\text{ISFA}}(\phi) &= a_{20} + c_{24}\cos(4\phi) + s_{24}\sin(4\phi) && (35)\\ \text{Case 2: } \Psi^{\mu\nu}_{\text{ISFA}}(\phi) &= a_{20} + c_{22}\cos(2\phi) + s_{22}\sin(2\phi) + \\ &\quad c_{24}\cos(4\phi) + s_{24}\sin(4\phi) && (36)\end{aligned}$$

with constants defined in Table 4. In the next step these objective functions are further simplified by combining the sine term and cosine term in a single cosine term. This results in:

$$\begin{aligned}\text{Case 1: } \Psi^{\mu\nu}_{\text{ISFA}}(\phi) &= A_0 + A_4\cos(4\phi + \phi_4) && (37)\\ \text{Case 2: } \Psi^{\mu\nu}_{\text{ISFA}}(\phi) &= A_0 + A_2\cos(2\phi + \phi_2) + A_4\cos(4\phi + \phi_4) && (38)\end{aligned}$$

with constants defined in Table 5. It is easy to see why it is possible to write both objective functions (37) and (38) in such a simple form. Firstly, the terms in (25) are products of at most four $\sin(\phi)$ and $\cos(\phi)$ functions, which allows, at most, a frequency of 4. Secondly, in Case 1 $\Psi^{\mu\nu}_{\text{ISFA}}(\phi)$ has a periodicity of $\pi/2$ because rotations by multiples of $\pi/2$ correspond to a permutation (possibly plus sign change) of the two components. Since both components are inside the subspace, permutations do not change the objective function and the objective function has a $\pi/2$ periodicity. Thus we conclude that only frequencies of 0 and 4 can be present in (37). In Case 2, since one component lies outside the subspace, an exchange of components will change the objective function (38). A rotation by multiples of $\pi$, however, which results only in a possible sign change, will leave the objective function unchanged, resulting in an objective function with $\pi$-periodicity and therefore frequencies of 0, 2, and 4.

[1] Freely available at http://mdp-toolkit.sourceforge.net .

|  | Case 1 | Case 2 |
|---|---|---|
| $d_0$ | $\sum\limits_{\tau\in T_{\mathrm{SFA}}} \kappa^{\tau}_{\mathrm{SFA}} \left(C^{(\mathbf{u}')}_{\mu\mu}\right)^2 + \left(C^{(\mathbf{u}')}_{\nu\nu}\right)^2$ | $\sum\limits_{\tau\in T_{\mathrm{SFA}}} \kappa^{\tau}_{\mathrm{SFA}} \left(C^{(\mathbf{u}')}_{\mu\mu}\right)^2$ |
| $d_1$ | $4 \sum\limits_{\tau\in T_{\mathrm{SFA}}} \kappa^{\tau}_{\mathrm{SFA}} \left(C^{(\mathbf{u}')}_{\mu\mu} C^{(\mathbf{u}')}_{\mu\nu} - C^{(\mathbf{u}')}_{\mu\nu} C^{(\mathbf{u}')}_{\nu\nu}\right)$ | $4 \sum\limits_{\tau\in T_{\mathrm{SFA}}} \kappa^{\tau}_{\mathrm{SFA}} C^{(\mathbf{u}')}_{\mu\nu} C^{(\mathbf{u}')}_{\mu\mu}$ |
| $d_2$ | $2 \sum\limits_{\tau\in T_{\mathrm{SFA}}} \kappa^{\tau}_{\mathrm{SFA}} \left(2\left(C^{(\mathbf{u}')}_{\mu\nu}\right)^2 + C^{(\mathbf{u}')}_{\mu\mu} C^{(\mathbf{u}')}_{\nu\nu}\right)$ | $2 \sum\limits_{\tau\in T_{\mathrm{SFA}}} \kappa^{\tau}_{\mathrm{SFA}} \left(2\left(C^{(\mathbf{u}')}_{\mu\nu}\right)^2 + C^{(\mathbf{u}')}_{\mu\mu} C^{(\mathbf{u}')}_{\nu\nu}\right)$ |
| $d_3$ | 0 | $4 \sum\limits_{\tau\in T_{\mathrm{SFA}}} \kappa^{\tau}_{\mathrm{SFA}} C^{(\mathbf{u}')}_{\mu\nu} C^{(\mathbf{u}')}_{\nu\nu}$ |
| $d_4$ | 0 | $\sum\limits_{\tau\in T_{\mathrm{SFA}}} \kappa^{\tau}_{\mathrm{SFA}} \left(C^{(\mathbf{u}')}_{\nu\nu}\right)^2$ |
| $d_c$ | $\sum\limits_{\tau\in T_{\mathrm{SFA}}} \kappa^{\tau}_{\mathrm{SFA}} \sum\limits_{\substack{\alpha=1,\\ \alpha\notin\{\mu,\nu\}}}^{R} \left(C^{(\mathbf{u}')}_{\alpha\alpha}\right)^2$ | $\sum\limits_{\tau\in T_{\mathrm{SFA}}} \kappa^{\tau}_{\mathrm{SFA}} \sum\limits_{\substack{\alpha=1\\ \alpha\neq\mu}}^{R} \left(C^{(\mathbf{u}')}_{\alpha\alpha}\right)^2$ |
| $e_0$ | $2 \sum\limits_{\tau\in T_{\mathrm{ICA}}} \kappa^{\tau}_{\mathrm{ICA}} \left(C^{(\mathbf{u}')}_{\mu\nu}\right)^2$ | $2 \sum\limits_{\tau\in T_{\mathrm{ICA}}} \kappa^{\tau}_{\mathrm{ICA}} \sum\limits_{\substack{\alpha=1\\ \alpha\neq\mu}}^{R} \left(C^{(\mathbf{u}')}_{\mu\alpha}\right)^2$ |
| $e_1$ | $4 \sum\limits_{\tau\in T_{\mathrm{ICA}}} \kappa^{\tau}_{\mathrm{ICA}} \left(C^{(\mathbf{u}')}_{\mu\nu} C^{(\mathbf{u}')}_{\nu\nu} - C^{(\mathbf{u}')}_{\mu\mu} C^{(\mathbf{u}')}_{\mu\nu}\right)$ | $\sum\limits_{\tau\in T_{\mathrm{ICA}}} \kappa^{\tau}_{\mathrm{ICA}} \sum\limits_{\substack{\alpha=1\\ \alpha\neq\mu}}^{R} C^{(\mathbf{u}')}_{\mu\alpha} C^{(\mathbf{u}')}_{\alpha\nu}$ |
| $e_2$ | $\sum\limits_{\tau\in T_{\mathrm{ICA}}} \kappa^{\tau}_{\mathrm{ICA}} \left(C^{(\mathbf{u}')}_{\mu\mu} - C^{(\mathbf{u}')}_{\nu\nu}\right)^2 - 2\left(C^{(\mathbf{u}')}_{\mu\nu}\right)^2$ | $2 \sum\limits_{\tau\in T_{\mathrm{ICA}}} \kappa^{\tau}_{\mathrm{ICA}} \sum\limits_{\substack{\alpha=1\\ \alpha\neq\mu}}^{R} \left(C^{(\mathbf{u}')}_{\alpha\nu}\right)^2$ |
| $e_c$ | $2 \sum\limits_{\tau\in T_{\mathrm{ICA}}} \kappa^{\tau}_{\mathrm{ICA}} \left(\sum\limits_{\alpha=1}^{R-1} \sum\limits_{\beta>\alpha}^{R} \left(C^{(\mathbf{u}')}_{\alpha\beta}\right)^2 - \left(C^{(\mathbf{u}')}_{\mu\nu}\right)^2\right)$ | $2 \sum\limits_{\tau\in T_{\mathrm{ICA}}} \kappa^{\tau}_{\mathrm{ICA}} \sum\limits_{\substack{\alpha=1,\\ \alpha\neq\mu}}^{R-1} \sum\limits_{\substack{\beta=\alpha+1,\\ \beta\neq\mu}}^{R} \left(C^{(\mathbf{u}')}_{\alpha\beta}\right)^2$ |

Table 3: Constants in Equation (25).

|  | Case 1 | Case 2 |
|---|---|---|
| $a_{20}$ | $\frac{b_{\mathrm{ICA}}}{4}(4e_c + e_2 + 3e_0)$<br>$- \frac{b_{\mathrm{SFA}}}{4}(4d_c + d_2 + 3d_0)$ | $\frac{b_{\mathrm{ICA}}}{2}(2e_c + e_0 + e_2)$<br>$- \frac{b_{\mathrm{SFA}}}{8}(8d_c + 3d_0 + d_2 + 3d_4)$ |
| $c_{22}$ | - | $\frac{b_{\mathrm{ICA}}}{2}(e_0 - e_2) - \frac{b_{\mathrm{SFA}}}{2}(d_0 - d_4)$ |
| $s_{22}$ | - | $\frac{b_{\mathrm{ICA}}}{2} e_1 - \frac{b_{\mathrm{SFA}}}{4}(d_1 + d_3)$ |
| $c_{24}$ | $\frac{b_{\mathrm{ICA}}}{4}(e_0 - e_2) - \frac{b_{\mathrm{SFA}}}{4}(d_0 - d_2)$ | $- \frac{b_{\mathrm{SFA}}}{8}(d_0 - d_2 + d_4)$ |
| $s_{24}$ | $\frac{b_{\mathrm{ICA}}}{4} e_1 - \frac{b_{\mathrm{SFA}}}{4} d_1$ | $- \frac{b_{\mathrm{SFA}}}{8}(d_1 - d_3)$ |

Table 4: Constants in Equation (35) and (36) in terms of the constants of Table 3.

|  | Case 1 | Case 2 |
|---|---|---|
| $A_0$ | $a_{20}$ | $a_{20}$ |
| $A_2$ | - | $\sqrt{c_{22}^2+s_{22}^2}$ |
| $A_4$ | $\sqrt{c_{24}^2+s_{24}^2}$ | $\sqrt{c_{24}^2+s_{24}^2}$ |
| $\tan(\phi_2)$ | - | $-\frac{s_{22}}{c_{22}}$ |
| $\tan(\phi_4)$ | $-\frac{s_{24}}{c_{24}}$ | $-\frac{s_{24}}{c_{24}}$ |

Table 5: Constants in Equations (37) and (38) in terms of the constants of Table 4.

# References

Almeida, L. (2004). Linear and nonlinear ICA based on mutual information - the MISEP method. *Signal Processing*, 84(2):231–245. Special Issue on Independent Component Analysis and Beyond.

Amari, S., Cichocki, A., and Yang, H. (1995). Recurrent neural networks for blind separation of sources. In *Proc. of the Int. Symposium on Nonlinear Theory and its Applications (NOLTA-95)*, pages 37–42, Las Vegas, USA.

Babaie-Zadeh, M., Jutten, C., and Nayebi, K. (2002). A geometric approach for separating post-nonlinear mixtures. In *Proc. of the XI European Signal Processing Conference (EUSIPCO 2002)*, pages 11–14.

Belouchrani, A., Abed Meraim, K., Cardoso, J.-F., and Éric Moulines (1997). A blind source separation technique based on second order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–44.

Blaschke, T., Berkes, P., and Wiskott, L. (2006). What is the relation between independent component analysis and slow feature analysis? *Neural Computation*, 18(10):2495–2508.

Blaschke, T. and Wiskott, L. (2004). CuBICA: Independent component analysis by simultaneous third- and fourth-order cumulant diagonalization. *IEEE Transactions on Signal Processing*, 52(5):1250–1256.

Cardoso, J.-F. (2001). The three easy routes to independent component analysis; contrasts and geometry. In *Proc. of the 3rd Int. Conference on Independent Component Analysis and Blind Source Separation, San Diego, (ICA 2001)*.

Cardoso, J.-F. and Souloumiac, A. (1996). Jacobi angles for simultaneous diagonalization. *SIAM J. Mat. Anal. Appl.*, 17(1):161–164.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314. Special Issue on Higher-Order Statistics.

Harmeling, S., Ziehe, A., Kawanabe, M., and Müller, K.-R. (2003). Kernel-based nonlinear blind source separation. *Neural Computation*, 15:1089–1124.

Hosseini, S. and Jutten, C. (2003). On the separability of nonlinear mixtures of temporally correlated sources. *IEEE Signal Processing Letters*, 10(2):43–46.

Hyvärinen, A. and Pajunen, P. (1999). Nonlinear independent component analysis: existence and uniqueness results. *Neural Networks*, 12(3):429–439.

Jutten, C. and Karhunen, J. (2003). Advances in nonlinear blind source separation. In *Proc. of the 4th Int. Symposium on Independent Component Analysis and Blind Signal Separation, Nara, Japan, (ICA 2003)*, pages 245–256.

McCullagh, P. (1987). *Tensor methods in statistics*. Monographs on Statistics and Applied Probability. Chapmann and Hall, London.

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

Taleb, A. (2002). A generic framework for blind source separation in structured nonlinear models. *IEEE Transactions on Signal Processing*, 50(8):1819–1830.

Taleb, A. and Jutten, C. (1997). Nonlinear source separation: The post-nonlinear mixtures. In *Proc. European Symposium on Artificial Neural Networks, Bruges, Belgium*, pages 279–284.

Taleb, A. and Jutten, C. (1999). Source separation in post non linear mixtures. *IEEE Transactions on Signal Processing*, 47(10):2807–2820.

Tong, L., Liu, R., Soon, V. C., and Huang, Y.-F. (1991). Indeterminacy and identifiability of blind identification. *IEEE Transactions on Circuits and Systems*, 38(5):499–509.

Wiskott, L. and Sejnowski, T. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770.

Yang, H.-H., Amari, S., and Cichocki, A. (1998). Information-theoretic approach to blind separation of sources in non-linear mixture. *Signal Processing*, 64(3):291–300.

Ziehe, A., Kawanabe, M., Harmeling, S., and Müller, K.-R. (2003). Blind separation of post-nonlinear mixtures using linearizing transformations and temporal decorrelation. *Journal of Machine Learning Research*, 4:1319–1338.

Ziehe, A. and Müller, K.-R. (1998). TDSEP – an efficient algorithm for blind separation using time structure. In *Proc. of the 8th Int. Conference on Artificial Neural Networks (ICANN'98)*, pages 675 – 680, Berlin. Springer Verlag.