Independent Subspace Analysis Using Geodesic Spanning Trees

Barnabás Póczos PBARN@CS.ELTE.HU
András Lőrincz ANDRAS.LORINCZ@ELTE.HU
Department of Information Systems, Eötvös Loránd University, 1117 Pázmány P. sétány 1/C, Budapest, Hungary
Research Group on Intelligent Information Systems, Hungarian Academy of Sciences, Budapest, Hungary

Abstract

A novel algorithm for performing Independent Subspace Analysis, the estimation of hidden independent subspaces is introduced. This task is a generalization of Independent Component Analysis. The algorithm works by estimating the multi-dimensional differential entropy. The estimation utilizes minimal geodesic spanning trees matched to the sample points. Numerical studies include (i) illustrative examples, (ii) a generalization of the cocktail-party problem to songs played by bands, and (iii) an example on mixed independent subspaces, where subspaces have dependent sources, which are pairwise independent.

1. Introduction

Independent Component Analysis (ICA) (Jutten & Herault, 1991; Comon, 1994) aims to recover linearly or non-linearly mixed independent, possibly noisy sources. There is broad range of applications for ICA, such as blind source separation and blind source deconvolution (Bell & Sejnowski, 1995), feature extraction (Bell & Sejnowski, 1997), denoising (Hyvärinen, 1999). Particular applications include, for example, the analysis of financial data (Kiviluoto & Oja, 1998), data from neurobiology, fMRI, EEG, and MEG (see, e.g., (Makeig et al., 1996; Vigário et al., 1998) and references therein). For a recent review on ICA see (Hyvärinen et al., 2001).

Original ICA algorithms are 1-dimensional in the sense that all sources are assumed to be independent, real valued stochastic variables. However, applications where not all sources, but groups of the sources are independent may be relevant in practice. In this case, independent sources can be multi-dimensional. For example, generalizations of the cocktail-party problem arise for independent groups of people talking about independent topics, or if groups of musicians play at the party. This is the Independent Subspace Analysis (ISA) extension of ICA, also called Multi-dimensional ICA (MICA) (Cardoso, 1998; Hyvärinen & Hoyer, 2000). An important applications is, e.g., the processing of EEG-fMRI data (Akaho et al., 1999). Efforts have been made to develop ISA algorithms (Cardoso, 1998; Vollgraf & Obermayer, 2001; Akaho et al., 1999; Bach & Jordan, 2003), but there are certain concerns with regard to these algorithms. Certain approaches use 2-dimensional Edgeworth-expansion (Akaho et al., 1999). This approach leads to sophisticated equations and it is not having higher dimensional generalizations yet. Another suggestion uses ICA as preprocessing step followed by permutations of the columns of the mixing matrix (Cardoso, 1998), but the method to find the right permutations has not been worked out. Another recent approach searches for independent subspaces via kernel methods (Bach & Jordan, 2003).

Here, we show that ISA needs the minimization of the sum of multi-dimensional differential entropies of the components that we estimate by means of minimal geodesic spanning trees (Hero & Michel, 1998; Yukich, 1998). The paper is organized as follows: In sections 2 and 3 the ISA model and the ISA cost function will be introduced. Details of entropy estimations are given in Section 4. Section 5 is on Jacobi-rotations that minimize the cost function. Numerical simulations are presented in Section 6. Results are discussed and conclusions are drawn in Section 7.

2. The ISA Model

Assume we have d of m-dimensional independent sources denoted by $\mathbf{y}^1, \ldots, \mathbf{y}^d$, respectively, where $\mathbf{y}^i \in \mathbf{R}^m$. Let $\mathbf{y} = [(\mathbf{y}^1)^T, \ldots, (\mathbf{y}^d)^T]^T \in \mathbf{R}^{dm}$, where superscript T stands for transposition. We assume that

Appearing in *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 2005.

these sources are hidden and we can observe only the following signals

$$\mathbf{x} = \mathbf{A}\mathbf{y} \tag{1}$$

where $\mathbf{A} \in \mathbf{R}^{dm \times dm}$. The task is to recover hidden source $\mathbf{y}$ and mixing matrix $\mathbf{A}$ given the observed signals $\mathbf{x} \in \mathbf{R}^{dm}$. In the ISA model we assume that $\mathbf{y}^i \in \mathbf{R}^m$ is independent of $\mathbf{y}^j \in \mathbf{R}^m$ for $i \neq j$. For the special case of m=1, the ICA problem is received.

In the ICA problem, given signals $\mathbf{x}$, sources y^i can be recovered only up to sign, up to arbitrary scaling factors and up to an arbitrary permutation. The ISA task has even more freedom, signals $\mathbf{y}^i$ can be recovered up to an arbitrary permutation and an m-dimensional linear, invertible transformation. It is easy to see this by considering matrix $\mathbf{C} \in \mathbf{R}^{dm \times dm}$ made of a permutation matrix of size $d \times d$ with each element made of an $m \times m$ block-matrix with invertible $\mathbf{C}_i$ blocks placed only to the non-zero elements of the permutation matrix. Then, $\mathbf{x} = \mathbf{A}\mathbf{y} = \mathbf{A}\mathbf{C}^{-1}\mathbf{C}\mathbf{y}$, and because $\mathbf{y}^i$ is independent of $\mathbf{y}^j$, thus $\mathbf{C}_i\mathbf{y}^i$ is independent of $\mathbf{C}_j\mathbf{y}^j$ $\forall i \neq j$. That is, in the ISA model, matrices $\mathbf{A}$ and $\mathbf{A}\mathbf{C}^{-1}$ and sources $\mathbf{y}^i$ and $\mathbf{C}_i\mathbf{y}^i$ are indistinguishable. The ambiguity of this task can be lowered by assuming

$$E\mathbf{y} = \mathbf{0}, \text{ and } E\{\mathbf{y}\mathbf{y}^T\} = \mathbf{I}_{md} \tag{2}$$

where E is the expected value operator, $\mathbf{I}_n$ is the n-dimensional identity matrix. Similarly, scaling of observed signals $\mathbf{x}$ can assure that

$$E\mathbf{x} = \mathbf{0}, \text{ and } E\{\mathbf{x}\mathbf{x}^T\} = \mathbf{I}_{md} \tag{3}$$

which is called the whitening of the inputs. Then, Eq. (1) ensures that $E\{\mathbf{x}\mathbf{x}^T\} = \mathbf{A}E\{\mathbf{y}\mathbf{y}^T\}\mathbf{A}^T$ and $\mathbf{I}_{md} = \mathbf{A}\mathbf{A}^T$. It then follows that signals $\mathbf{y}^i$ can be recovered up to *permutation* and up to m-dimensional *orthogonal transformation* in the ISA problem. In other words, if $\mathbf{C}_i \in \mathbf{R}^{m \times m}$ is an orthogonal matrix, then signals $\mathbf{x}$ will not provide information if the original sources correspond to $\mathbf{y}^i$ or to $\mathbf{C}_i\mathbf{y}^i$. In 1D, this is equivalent to the uncertainty of the sign of y^i ($\mathbf{C}_i = 1$ or $\mathbf{C}_i = -1$). Thus, without loss of generality, we can restrict the search for mixing matrix $\mathbf{A}$ and separation matrix $\mathbf{W}$, to the set of orthogonal matrices.

3. The ISA Cost Function

We shall derive the cost function for ISA task under the constraint $\mathbf{W}^T\mathbf{W} = \mathbf{I}_{md}$ for the separation matrix $\mathbf{W}$. Global minima of this cost function will include separation matrix $\mathbf{W}$. Let us introduce the following notations. Let $I(\mathbf{y}^1, \ldots, \mathbf{y}^d)$ denote the mutual information between vectors $\mathbf{y}^1, \ldots, \mathbf{y}^d \in \mathbf{R}^m$:

$$I(\mathbf{y}^1, \ldots, \mathbf{y}^d) = \int p(\mathbf{y}) \log \frac{p(\mathbf{y})}{p(\mathbf{y}^1) \ldots p(\mathbf{y}^d)} d\mathbf{y} \tag{4}$$

where $p(\mathbf{y}) = p(\mathbf{y}^1, \ldots, \mathbf{y}^d)$ denotes the joint probability density function of stochastic variables $\mathbf{y}^1, \ldots, \mathbf{y}^d$ and $p(\mathbf{y}^j)$ denotes the marginal density of $\mathbf{y}^j$.

Now, for a real valued stochastic vector variable $\mathbf{y}$ let the differential entropy be denoted by $H(\mathbf{y})$, that is

$$H(\mathbf{y}) = -\int p(\mathbf{y}) \log p(\mathbf{y}) d\mathbf{y} \tag{5}$$

Further, assume $\mathbf{y} = \mathbf{W}\mathbf{x}$. Then

$$I(\mathbf{y}^1, \ldots, \mathbf{y}^d) = -H(\mathbf{x}) + \log|\mathbf{W}| + \sum_{i=1}^{d} H(\mathbf{y}^i) \tag{6}$$

because $H(\mathbf{W}\mathbf{x}) = H(\mathbf{x}) + \log|\mathbf{W}|$. However, $H(\mathbf{x})$ is constant and $\log|\mathbf{W}| = 0$ since $\mathbf{W}^T\mathbf{W} = \mathbf{I}$. Thus, our task is to minimize

$$J(\mathbf{W}) \doteq H(\mathbf{y}^1) + \ldots + H(\mathbf{y}^d) \tag{7}$$

that is, the solution of the ISA task is equivalent to the minimization of the multi-dimensional entropies of the marginals of the corresponding vector variables.

4. Multi-dimensional Entropy Estimation

Shannon's differential entropy of the $H(\mathbf{y}^i)$ terms of Eq. (7) needs to be estimated. First, we shall estimate Rényi's α-entropy of stochastic variable $\mathbf{y}$ having probability density function f, which is defined as

$$H_\alpha \doteq \frac{1}{1-\alpha} \log \int f^\alpha(\mathbf{y}) d\mathbf{y} \tag{8}$$

It is known that in the limit $\lim_{\alpha \to 1} H_\alpha = -\int f(\mathbf{y}) \log f(\mathbf{y}) d\mathbf{y}$ Rényi's entropy converges to Shannon's entropy. Rényi's entropy has already been used for ICA problems (Hild et al., 2001). To our best knowledge, it has not been used for the ISA task.

Let $\{\mathbf{y}^i(1), \ldots, \mathbf{y}^i(n)\}$ denote n independent and identically distributed (i.i.d.) samples drawn from distribution $\mathbf{y}^i \in \mathbf{R}^m$. The γ *weighted Euclidean graph* belonging to these points is a spanning tree of the points having edges $E = \{\mathbf{e} : \mathbf{e} = \mathbf{y}^i(p) - \mathbf{y}^i(q) \in \mathbf{R}^m, p \neq q\}$ between the points and having edge lengths equal to the γ^{th} power of the Euclidean distance ($l_\gamma(\mathbf{e}) = \|\mathbf{e}\|^\gamma$). A graph is called *minimal (γ weighted) Euclidean graph*, if the sum of the edge lengths of the graph is minimal, i.e.,

$$L_\gamma(\mathbf{y}^i) = \min_{T \in \mathcal{T}} \sum_{e \in T} \|\mathbf{e}\|^\gamma \tag{9}$$

where $\mathcal{T}$ denotes the set of spanning trees belonging to node set $\{\mathbf{y}^i(1), \ldots, \mathbf{y}^i(n)\}$. Let $\gamma = m - m\alpha$.

With our notations, and under certain technical conditions, the Beadword-Halton-Hammersley Theorem holds (Yukich, 1998), and one can show that

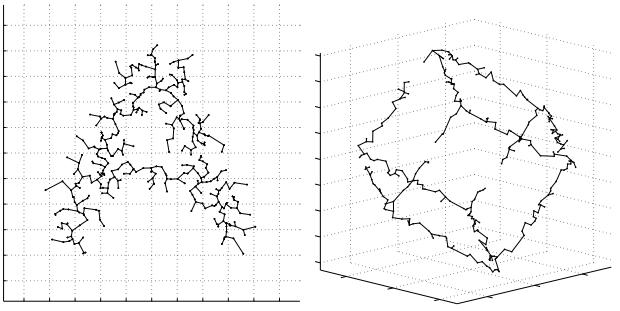
$$\frac{m}{\gamma}\log\frac{L_\gamma(\mathbf{y}^i)}{n^\alpha}\to H_\alpha(\mathbf{y}^i)+c,\text{ as } n\to\infty \qquad (10)$$

with constant c being independent of the distribution of $\mathbf{y}^i$. This estimation is asymptotically unbiased and strongly consistent (Yukich, 1998). However, this estimation is sensitive to outliers if the spanning tree has long edges. Different methods have been derived to increase the robustness of the estimation.

One approach deletes the longest k edges from the spanning tree and considers the remaining sum of the edge lengths (Banks et al., 1992). Another approach replaces the set of points $\mathbf{y}^i(1),\ldots,\mathbf{y}^i(n)$ by the k-element subset which produces the shortest sum of edge lengths. Although this task is NP-complete, it seems to have an effective greedy approximation (Hero & Michel, 1998).

We shall follow a third route, which utilizes *geodesic distances*. This method uses the edges of the k-nearest nodes or the nodes within radius ε to each node $\mathbf{y}^i(p)$. These sub-graphs are called the *Euclidean neighborhood graph*. One needs to find the minimal spanning forest made of such subgraphs called the *geodesic spanning forest* of the set. The geodesic distance between two points is defined as the length of the shortest path on the geodesic spanning forest. These methods are used in manifold learning problems, such as ISOMAP (Tenenbaum et al., 2000) as well as for the estimation of intrinsic dimensions (Costa & Hero, 2004). Neighborhood graphs with neighbors $k=10$ were constructed, the geodesic spanning forests were computed for illustrative purposes: Figure 1(a) and 1(b) show the geodesic spanning forests for letter 'A' sampled uniformly and that of a three dimensional cubic wireframe sampled uniformly and with added noise of 2% standard deviation, respectively.

The limit $\alpha\to 1$ in Eq. (10) leads to an estimation of Shannon's entropy. Trivially, $\alpha\to 1$ exactly when $\gamma\to 0$. In practice, limit $n\to\infty$ can only be approximated. Computation of the limit $\gamma\to 0$ is also troublesome and small but fixed γ values are used. Numerical experiments demonstrate the quality of this entropy estimation (Fig. 2): We draw 10000 samples from two of 3 dimensional independent normal distributions, each having randomly chosen non-diagonal covariance matrices. The samples were then mixed by rotation matrix $G(\theta)\in\mathbf{R}^{(6\times 6)}$, where parameter θ were chosen from interval $[-\pi,\pi]$. Figure 2 shows the exact and the estimated entropies of the first 3 coordinates of the mixed vectors. The two curves differ only by and additive and irrelevant constant.



(a) 2D example for letter 'A' (b) 3D example for the edges of a cube

Figure 1. **Examples for geodesic spanning forests.** Forests for (a) 400 samples taken from letter 'A' with line thickness of 15% of the letter size and (b) 400 samples of a noisy 3D cubic wireframe.

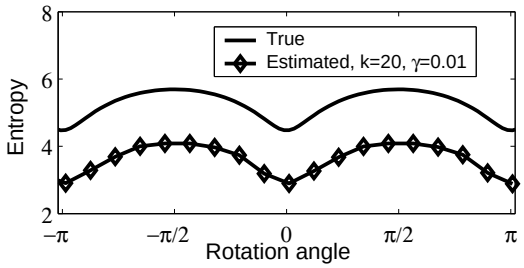


Figure 2. **Estimated and true entropies.** Solid line: entropy computed analytically for a 3 dimensional problem. Solid line with diamonds: estimated entropy using Eq. (10) for $\gamma=0.01$ and $k=20$. The two curves differ by an (irrelevant) additive constant.

Computation of the Euclidean neighborhood graph needs $O(n^2)$ steps, where n is the number of samples. In our case, the minimal geodesic spanning forest can be found in $O(n\log n)$ steps with Kruskal's method. That is, the computational costs of the entropy estimation in Eq. (10) scales as $O(n^2)$.

5. Optimization

First, we shall discuss how to optimize the cost function (7) by means of Jacobi-rotations. Then the pseudo-code of our algorithm will be provided. Finally, we shall generalize the Amari-distance (Amari et al., 1996) to provide a performance measure for ISA tasks.

Let us rewrite Eq. (7) in a different form by taking advantage of relation $H(\mathbf{y}^j)=\sum_{i=1}^m H(y_i^j)-$

$I(y_1^j, \ldots, y_m^j)$. Thus we have

$$J(\mathbf{W}) = \sum_{j=1}^{d} \sum_{i=1}^{m} H(y_i^j) - \sum_{j=1}^{d} I(y_1^j, \ldots, y_m^j) \qquad (11)$$

One may proceed in minimizing $J(\mathbf{W})$ by first estimating the different terms of the cost function for a given $\mathbf{W}$, that is the multi-dimensional $H(\mathbf{y}^j)$ entropies in Eq. (7) or the one dimensional $H(y_i^j)$ entropies and the mutual information $I(y_1^j, \ldots, y_m^j)$ in Eq. (11) and then minimizing these quantities by changing $\mathbf{W}$. Considering Eq. (11), a two-step heuristics that provides local minima can be derived. Note that

$$\min_{\mathbf{W}} \sum_{j=1}^{d} \sum_{i=1}^{m} H(y_i^j) \qquad (12)$$

is equivalent to the classical 1D ICA problem (Learned-Miller & Fisher, 2003) and in the *first step*, one can solve the traditional 1D ICA task, which minimizes the expression $\sum_{j=1}^{d} \sum_{i=1}^{m} H(y_i^j)$. In the second step, the quantity $\sum_{j=1}^{d} I(y_1^j, \ldots, y_m^j)$, which is the sum of the mutual information of the subspaces should be maximized. Although there are methods for the estimation of the mutual information (Bach & Jordan, 2002; Gretton et al., 2003), we have found that good results may be achieved by estimating only the multi-dimensional entropy in Eq. (7) and using a simple heuristics, an improvement of the permutation method mentioned in (Cardoso, 1998). We note that no permutation of the ICA components will modify the term $\sum_{j=1}^{d} \sum_{i=1}^{m} H(y_i^j)$ and such permutations will improve the cost function Eq. (11) if they increase the quantity $\sum_{j=1}^{d} I(y_1^j, \ldots, y_m^j)$. The idea is to group components that belong to each other. In the general case, however, the ICA components may need to be modified for proper grouping and the approximation needs to be extended. This extension is detailed below.

5.1. Jacobi-Rotations

Let $\mathbf{W}^* \in \mathbf{R}^{md \times md}$ denote an optimum of $J(\mathbf{W})$. One does not need to explore the full $\mathbf{R}^{md \times md}$ space for finding an optimum, because $\mathbf{W}$ is orthogonal and thus the space to be searched is somewhat smaller; it has $md(md-1)/2$ dimensions. Now, let us consider Jacobi-rotation (also known as Givens-rotation). Let $1 \leq p < q \leq md$, and let θ denote a rotation angle. Jacobi-rotation of angle θ of components p and q is denoted by $\mathbf{G}(p, q, \theta) \in \mathbf{R}^{md \times md}$. This matrix is derived from identity matrix $\mathbf{I}_{md}$ by modifying 4 of its elements: $\mathbf{G}(p, q, \theta)_{pp} = \mathbf{G}(p, q, \theta)_{qq} = \cos(\theta)$, $\mathbf{G}(p, q, \theta)_{qp} = -\mathbf{G}(p, q, \theta)_{pq} = \sin(\theta)$. Thus $\mathbf{G}$ is orthogonal and for any vector $\mathbf{y} \in \mathbf{R}^{md}$ rotation by matrix $\mathbf{G}(p, q, \theta)$ mixes only the p^{th} and the q^{th} elements of vector $\mathbf{y}$. In our approach we shall avoid the optimization of the orthogonal matrix in the full $\mathbf{R}^{md \times md}$ space, because such global optimization is intractable. Instead, we shall perform a series of 1D optimization characterized by angle θ. Note that there is no need to consider all pairs $1 \leq p < q \leq md$ in the optimization. Optimization may be restricted to those p, q pairs that belong to different subspaces.

In the *second step* we shall execute a series of iteration cycles. One cycle of the iteration has $m^2 d(d-1)/2$ separate 1D Jacobi-rotation optimization tasks. The rotations will either demix two components ($0 < |\theta| < \pi/2$), or may leave those unchanged ($\theta = 0$), or may exchange them ($|\theta| = \pi/2$). Iteration will stop if new cycles can not improve the results. We note that global minimum may not be reached by this algorithm. However, random rotations of the subspaces between iteration cycles diminishes the chances of being trapped in local minima, provided that simulated annealing strategy is applied. Such random rotations were not used here.

5.2. Algorithm

Traditional ICA preprocessing on the data was followed by a series of Jacobi-rotations. For all rotations the global minimum of cost function J was computed by single dimensional ($\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$) exhaustive search, where entropy was estimated through Eq. (10). The pseudo-code is provided below:

```
INITALIZATION
   Sources:  number d, dimension m
   x(1), ..., x(n) ∈ R^md: measured signals
PREPROCESSING
   y ∈ R^md:  ICA estimations from x
MAIN PROGRAM
Iterate until convergence
   for p = 1 : md − m,  q = m⌊(p−1)/m⌋ + m + 1 : dm

      Choose θ* := arg max_{θ∈[−π/2, π/2]} J_{p,q}(θ)
         where
         J_{p,q}(θ) := H(ŷ^1(θ)) + ... + H(ŷ^d(θ)),
         [ŷ^1(θ)^T, ..., ŷ^d(θ)^T]^T := G(p, q, θ)y,
         and
         H is computed by Eq. (10).
      Set y := G(p, q, θ*)y

   endfor
OUTPUT
Estimated sources:  y* := y
```

5.3. Generalized Amari-Distance

We know for ICA that the product of a good estimation for separation matrix $\mathbf{W}$ and the original mixing matrix $\mathbf{A}$ is a permutation matrix. Deviation from the permutation matrix is measured by the Amari-distance (Amari et al., 1996). This measure is widely used for performance estimation of ICA algorithms. Let $\mathbf{WA} = \mathbf{B}$. Then the Amari-distance $\rho(\mathbf{A}, \mathbf{W})$ of the estimated separation matrix $\mathbf{W}$ and mixing matrix $\mathbf{A}$ is defined as:

$$\rho(\mathbf{A}, \mathbf{W}) = \frac{1}{2d} \sum_{i=1}^{d} \left(\frac{\sum_{j=1}^{d} |b_{ij}|}{\max_j |b_{ij}|} - 1 \right) + \\ + \frac{1}{2d} \sum_{j=1}^{d} \left(\frac{\sum_{i=1}^{d} |b_{ij}|}{\max_i |b_{ij}|} - 1 \right) \qquad (13)$$

Multi-dimensional generalization of the Amari-distance is straightforward. Here, matrix $\mathbf{WA} \in \mathbf{R}^{md \times md}$ is a permutation matrix permuting $m \times m$ block matrices. Deviation from this matrix can be computed as follows. Let b_{ij} denote the sum of the absolute values of elements at the intersection of the $i(m-1)+1, \ldots, im$ rows and the $j(m-1)+1, \ldots, jm$ columns of matrix $\mathbf{WA}$. Formally, let $\mathbf{C} = \mathbf{WA}$ and for $1 \leq i, j \leq d$ let

$$b_{ij} = \sum_{p=i(m-1)+1}^{im} \sum_{q=j(m-1)+1}^{jm} |C_{pq}| \qquad (14)$$

It is easy to see that $\rho(\mathbf{A}, \mathbf{W}) \geq 0$ and it is zero if and only if matrix $\mathbf{WA}$ is a permutation matrix permuting $m \times m$ block matrices.

6. Numerical Simulations

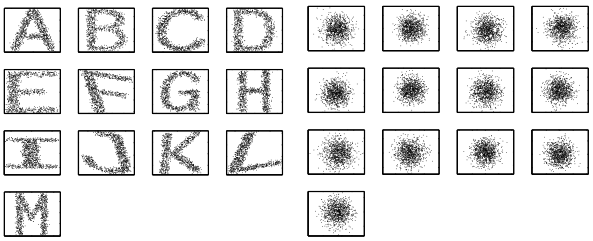
In this section numerical simulations will be used to demonstrate the convergence of the ISA algorithm.

6.1. Simple 2D Letters and 3D Curves

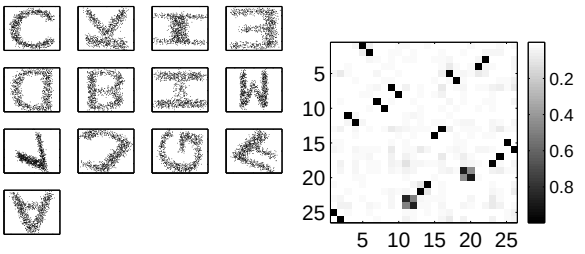
13 (6) pieces of 2 (3) dimensional independent sources were chosen, none of them were linearly separable in 2 (3) dimensional spaces. For the sake of visualization, sources formed simple 2 (3) dimensional patterns. First, the points on these patterns were sampled independently. 2D samples were generated from letters, alike in Fig. 1(a). For the 3D case, samples were generated from noise-free 3D wirefames. The generated sample points (2000 in number) were whitened. These 2 (3) dimensional sources are shown in Fig. 3(a) (Fig. 3(e)). Random matrix of dimension 26×26 (18×18) was used to mix the sources. 13 (6) pieces of 2 (3) dimensional projections of the mixed sources are shown in Fig. 3(b) (Fig. 3(f)). The ISA algorithm was applied to the mixed signals with $k = 20$ neighbors and with γ value equal to 0.01. Results of the separation are shown in Fig. 3(c) (Fig. 3(g)). The ISA algorithm could recover the sources up to permutation and the directions within the subspaces. This feature is illustrated in Fig. 3(d) (Fig. 3(h)) by the performance matrix, which is the product of the true mixing matrix and the estimated separation matrix. This matrix, as expected, is close to a permutation matrix made of $m \times m$-sized blocks. Figs. 4(a) and 4(b) show the Amari distances (Eq. (13)) of the estimated matrices and the separation matrices during the course of the iterations for the 2D and 3D curves, respectively.

6.2. Non-linear Speakers

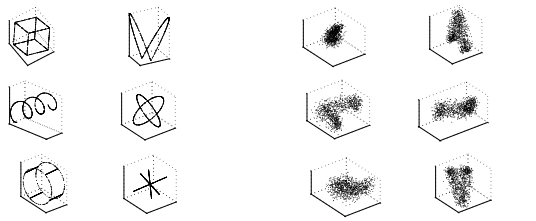
In the well known cocktail party problem n independent sources (speakers) and n microphones are assumed in a room. The task is to recover the original independent sources from the mixed signals received by the microphones. Here we modify the original problem. Assume d independent sources and to each source assume m pieces of non-linear speakers. The room also has md microphones. The task is to recover the original signals from the signals detected by the microphones. Traditional ICA algorithms can not handle this task, because the md sources are not independent; we have only d independent sources. The non-linear speakers distort the problem and to each source we have an m-dimensional linearly inseparable subspace. Clearly, for $m = 1$ the problem reduces to the original cocktail party problem. In the numerical studies $d = 4$ and $m = 2$ were used. For each source, one of the speakers had a cubic ($f(x) = x^3$) non-linearity. For the sake of visualization, points of the original signals are depicted as follows. Horizontal coordinate represents the signals emitted by the non-linear speaker, whereas vertical coordinate represents its counterpart emitted by the linear speaker. Figure 5(a) shows the 2-dimensional independent sources after whitening. The sources were mixed by a random matrix of size $md \times md$ modelling md microphones dispersed in the room. Projections of the mixed signals are depicted in Fig. 5(b). Mixed signals were then analyzed by algorithm (5.2) for neighbor number k=20 and for edge exponent $\gamma = 0.01$. Estimated sources are depicted in Fig. 5(c). The product of the original mixing matrix and the estimated separation matrix is shown in Fig. 5(d). This matrix closely approximates a permutation block matrix.



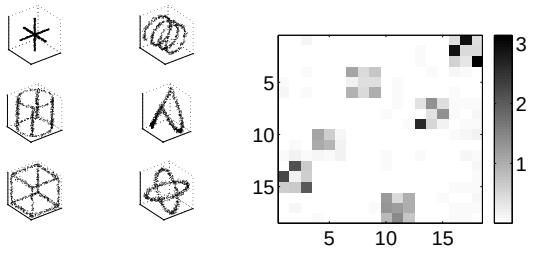
(a) 2D independent sources (b) 2D mixed sources



(c) 2D estimated sources (d) 2D performance matrix



(e) 3D independent sources (f) 3D mixed sources



(g) 3D estimated sources (h) 3D performance matrix

Figure 3. **ISA results for 2D and 3D curves**

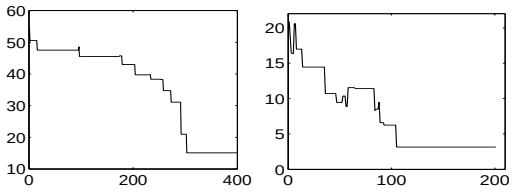
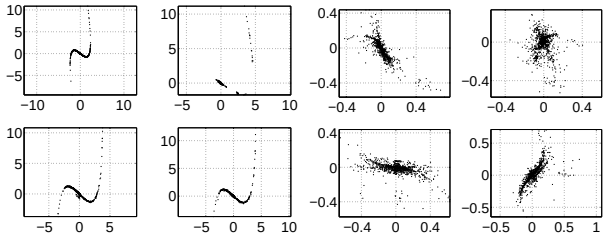
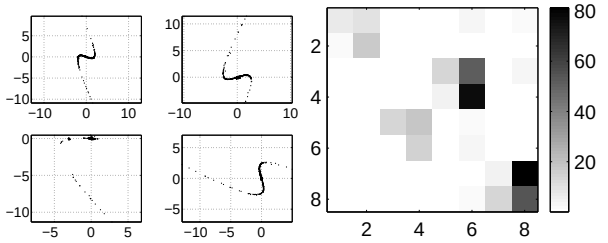


Figure 4. **Amari-distances versus iteration number.** Left: 2D, right: 3D.



(a) Whitened independent sources (b) Mixed sources



(c) Estimated sources (d) Product of mixing and separation matrices

Figure 5. **Non-linear speakers**

6.3. Separation of Beatles Songs

One can consider the straightforward generalization of the cocktail party problem, where the independent individual speakers are replaced by independent groups of speakers. We have taken 3 different Beatles songs modelling that three groups are playing. For each songs there are 4 sound tracks (4 sources). In order to strengthen the independence of the 3 songs, we have shuffled each source in time separately. Then the 12 sources were mixed by a random matrix of size 12×12. These mixed signals formed the inputs of the ISA algorithm. If there were a single sound track for each song then the task would reduce to the original cocktail party problem. The product of the original mixing matrix and the separation matrix estimated by ISA algorithm (5.2) is shown in Fig. 6. The product is close to a permutation block matrix permuting blocks of size 4×4. Thus our method, which can not recover the original individual tracks, is capable to recover the individual songs. The 1D ICA algorithms can not solve this generalized task. We note however, that temporal correlations are strong and ISA results are also poor without shuffling the different songs differently.

6.4. Groups of Pairwise Independent Sources

We note that in the ICA task pairwise independence of the variables and independence of all variables are equivalent under mild conditions. This is, however, not true for the ISA task as we shall see it below.

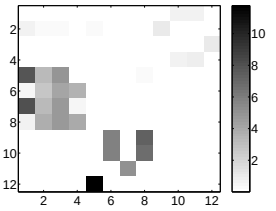


Figure 6. **Product of the mixing matrix and the estimated separation matrix for mixed Beatles songs**

Consider the following statement (Comon, 1994):

Lemma 6.1 *Let $\mathbf{x} \in \mathbf{R}^n$ be a stochastic vector variable with independent components having at most a single component of Gaussian probability distribution. Let the distribution functions of all other components continuous, too. Let us mix the components by orthogonal matrix $\mathbf{C} \in \mathbf{R}^{n \times n}$, that is let $\mathbf{z} = \mathbf{Cx}$. Then the following statements are equivalent:*

1. *Components z_i are pairwise independent*
2. *Components z_i are jointly independent*

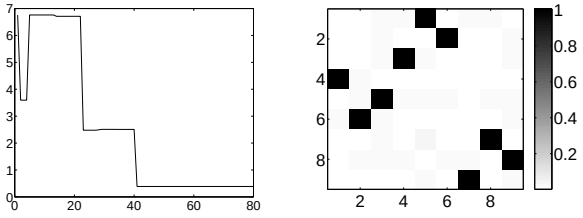
This theorem ensures that if we know that our data are mixed from jointly independent sources, then for the ICA task it is sufficient to require pairwise independence. Algorithms, which take advantage of this theorem include, for example the diagonalization of the non-linear covariance matrix (Jutten & Herault, 1991), or kernel-ICA algorithms (Bach & Jordan, 2002).

The statement however does not carry over to ISA tasks. Consider, for example, 3 groups of 3-dimensional sources $\{s_1^i, s_2^i, s_3^i\}$, $i = 1 \ldots 3$, which generate signals independently. Assume that for the i^{th} group, pairs s_j^i, s_k^i components are pairwise independent if $j \neq k$, but s_1^i, s_2^i, s_3^i are not independent. It is easy to see that if the independent groups are shuffled by permutation matrix $\mathbf{C} \in \mathbf{R}^{9 \times 9}$, then the new components, e.g.,

$$\{s_1^1, s_1^2, s_1^3\}, \{s_2^1, s_2^2, s_2^3\}, \{s_3^1, s_3^2, s_3^3\}$$

are pairwise independent but are not solutions to the ISA task, because they are not jointly independent. The efficiency of our ISA algorithm is demonstrated for this case below:

The following sources were prepared. 2 dices of 6 sides were thrown. Assume that the results are u_1 and u_2. Set the sources as follows: Let $s_1^1 = 1$ and $s_1^1 = -1$, if u_1 is even or odd, respectively. Let $s_2^1 = 1$ and $s_2^1 = -1$ if u_2 is odd or even, respectively. Further, let $s_3^1 = 1$ and $s_3^1 = -1$ if $u_1 - u_2$ is even or odd, respectively. It is easy to see that any two of sources s_1^1, s_2^1, s_3^1 are independent, but there are dependencies amongst them. For example, if $s_1^1 = s_2^1 = 1$, then we know that $s_3^1 = -1$. Slight modification can make the distribution $\mathbf{s}^1 = (s_1^1, s_2^1, s_3^1)$ continuous. We have used 3 of such 3D sources, mixed them by a random mixing matrix of size 9×9 and applied the ISA algorithm (5.2) with parameters k=20 and $\gamma = 0.01$. The algorithm recovered the original subspaces. This is demonstrated by the Amari error curve (Fig. 7(a)) and by the block permutation form of the performance matrix (Fig. 7(b)).



(a) Amari error during learning (b) Performance matrix

Figure 7. **Separation of blocks of dependent sources, which are pairwise independent**

7. Discussion and Conclusions

We have introduced a cost function (Eq. (7)) for estimating the solutions of ISA problems by searching for minimal costs. The cost function is the multi-dimensional generalization of the $\sum_i \sum_j H(y_j^i)$ 1D cost, frequently applied in ICA tasks. Equation (11), which is an equivalent form of Eq. (7), shows that it might be useful to start ISA optimization by using traditional ICA estimations. Also, the ISA task combines the traditional ICA task and a number of under-complete anti-ICA tasks, where anti-ICA means that mutual information is to be maximized instead of minimization.

The multi-dimensional entropy terms of the cost function were estimated by means of minimal geodesic spanning forests. During the minimization procedure, Jacobi-rotations were applied, which correspond to optimizations in single dimensions. In these one-dimensional optimizations the one-dimensional spaces were discretized and exhaustive searches were applied. The efficiency of the algorithm was demonstrated on a series of numerical examples.

Our method is relatively fast, because we need to do exhaustive searches only for the Jacobi-rotations, i.e., only in one dimensions. Our method estimates joint

dependencies and, in turn, it can overcome pitfalls that might arise under the restricted assumption of pairwise independence.

Our algorithm uses the information about the dimension of the sources and these dimensions were set equal in the simulations. Generalization to sources of non-equal dimensions is straightforward. The generalization to unknown dimensions is, however, non-trivial. Approximate solution has been suggested in (Bach & Jordan, 2003).

References

Akaho, S., Kiuchi, Y., & Umeyama, S. (1999). MICA: Multimodal independent component analysis. *Proc. IJCNN* (pp. 927–932).

Amari, S., Cichocki, A., & Yang, H. (1996). A new learning algorithm for blind source separation. *NIPS* (pp. 757–763).

Bach, F. R., & Jordan, M. I. (2002). Kernel independent component analysis. *JMLR*, *3*, 1–48.

Bach, F. R., & Jordan, M. I. (2003). Finding clusters in independent component analysis. *Fourth Int. Symp. on ICA and BSS* (pp. 891–896).

Banks, D., Lavine, M., & Newton, H. J. (1992). The minimal spanning tree for nonparametric regression and structure discovery. *Comput. Sci. and Stat.*, *24*, 370–374.

Bell, A. J., & Sejnowski, T. J. (1995). An information maximisation approach to blind separation and blind deconvolution. *Neural Comp.*, *7*, 1129–1159.

Bell, A. J., & Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. *Vision Research*, *37*, 3327–3338.

Cardoso, J. (1998). Multidimensional independent component analysis. *Proc. ICASSP'98, Seattle, WA.*

Comon, P. (1994). Independent component analysis, a new concept? *Signal Proc.*, *36*, 287–314.

Costa, J. A., & Hero, A. O. (2004). Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Trans. on Signal Proc.*, *52*, 2210–2221.

Gretton, A., Herbrich, R., & Smola, A. (2003). The kernel mutual information. *Proc. ICASSP.*

Hero, A., & Michel, O. (1998). Robust entropy estimation strategies based on edge weighted random graphs. *Proc. SPIE98* (pp. 250–261). San Diego, CA.

Hild, K. E., Erdogmus, D., & Príncipe, J. (2001). Blind source separation using Renyi's mutual information. *IEEE Signal Proc. Letters*, *8*, 174–176.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis.* New York: John Wiley.

Hyvärinen, A. (1999). Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation. *Neural Comp.*, *11*, 1739–1768.

Hyvärinen, A., & Hoyer, P. (2000). Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Comp.*, *12*, 1705–1720.

Jutten, C., & Herault, J. (1991). Blind separation of sources: An adaptive algorithm based on neuromimetic architecture. *Signal Proc.*, *24*, 1–10.

Kiviluoto, K., & Oja, E. (1998). Independent component analysis for parallel financial time series. *Proc. ICONIP'98* (pp. 895–898).

Learned-Miller, E. G., & Fisher, J. W. (2003). ICA using spacings estimates of entropy. *JMLR*, *4*, 1271–1295.

Makeig, S., Bell, A. J., Jung, T. P., & Sejnowski, T. J. (1996). Independent component analysis of electroencephalographic data. *NIPS* (pp. 145–151).

Tenenbaum, J. B., de Silva, V., & Langford., J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*, 626–634.

Vigário, R., Jousmaki, V., Hamalainen, M., Hari, R., & Oja, E. (1998). Independent component analysis for identification of artifacts in magnetoencephalographic recordings. *NIPS* (pp. 229–235).

Vollgraf, R., & Obermayer, K. (2001). Multidimensional ICA to separate correlated sources. *NIPS* (pp. 993–1000).

Yukich, J. E. (1998). *Probability theory of classical euclidean optimization problems*, vol. 1675 of *Lecture Notes in Math.* Springer-Verlag, Berlin.