ICA and ISA Using Schweizer-Wolff Measure of Dependence

Keywords: independent component analysis, independent subspace analysis, copula, non-parametric estimation of dependence

Abstract

We propose a new algorithm for independent component and independent subspace analysis problems. This algorithm uses a contrast based on the Schweizer-Wolff measure of pairwise dependence (Schweizer & Wolff, 1981), a non-parametric measure based on pairwise ranks of the variables. Our algorithm frequently outperforms state of the art ICA methods in the normal setting, is significantly more robust to outliers in the mixed signals, and performs well even in the presence of noise. Since pairwise dependence is evaluated explicitly, using Cardoso's conjecture (Cardoso, 1998), our method can be applied to solve independence subspace analysis (ISA) problems by grouping signals recovered by ICA methods. We provide an extensive empirical evaluation using simulated, sound, and image data.

1. Introduction

Independent component analysis (ICA) (Comon, 1994) deals with a problem of a blind source separation under the assumptions that the sources are independent and that they are linearly mixed. ICA has been used in the context of blind source separation and deconvolution, feature extraction, denoising, and successfully applied to many domains including finances, neurobiology, and processing of fMRI, EEG, and MEG data. For recent reviews on ICA see Hyvärinen et al. (2001).

Independent subspace analysis (ISA) (also called multi-dimensional ICA and group ICA) is a generalization of ICA that assumes that certain sources depend on each other, but the dependent groups of sources are still independent of each other, i.e., the independent groups are multidimensional. The ISA task has been the subject of extensive research (e.g., Cardoso, 1998; Theis, 2005; Bach & Jordan, 2003; Hyvärinen & Köster, 2006; Póczos & Lőrincz, 2005). and applied, for instance, to EEG-fMRI data (Akaho et al., 1999).

Our contribution, SWICA, is a new ICA algorithm based on Schweizer-Wolff (SW) non-parametric dependence measure. SWICA has the following properties:

- SWICA performs comparably to other state of the art ICA methods, outperforming them in a large number of test cases.
- SWICA is extremely robust to outliers as it uses *rank* values of the signals rather than their actual values.
- SWICA suffers less from the presence of noise than other algorithms.
- SW measure can be used as the cost function to solve ISA problems by grouping sources recovered by ICA methods.
- SWICA is simple to implement, and the Matlab/C++ code is available for public use.[1]
- On a negative side, SWICA is slower than other methods, limiting its use to sources of moderate dimensions, and it requires more samples to demix sources with near-Gaussian distributions.

The paper is organized as follows. An overview of the ICA and ISA problems and methods is presented in Section 2. Section 3 motivates and describes Schweizer-Wolf dependence measure. Section 4 describes a 2-source version of SWICA, extends it to a d-source problem, describes an application to ISA, and mentions possible approaches for accelerating SWICA. Section 5 provides a thorough empirical evaluation of

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1] Our implementation will become publicly available once the algorithm is accepted for publication.

SWICA to other ICA algorithms under different settings and data types. The paper is concluded with a summary in Section 6.

2. ICA and ISA

We consider the following problem. Assume we have d independent 1-dimensional sources (random variables) denoted by $S^1, \ldots, S^d$. We assume each source emits N i.i.d. samples denoted by $\left(s_1^i, \ldots, s_N^i\right)$. Let $\mathbf{S} = \left\{s_i^j\right\} \in \mathbb{R}^{d \times N}$ be a matrix of these samples. We assume that these sources are hidden, and that only a matrix $\mathbf{X}$ of mixed samples can be observed:

$$\mathbf{X} = \mathbf{AS}$$

where $\mathbf{A} \in \mathbb{R}^{d \times d}$. (We further assume that $\mathbf{A}$ has full rank d.) The task is to recover the sample matrix $\mathbf{S}$ of the hidden sources by finding a demixing matrix $\mathbf{W}$

$$\mathbf{Y} \quad = \quad \mathbf{WX} = (\mathbf{WA})\,\mathbf{S},$$

and the estimated sources $Y^1, \ldots, Y^d$ are mutually independent. The solution can be recovered only up to a scale and a permutation of the components, so the mixed signals are usually preprocessed (pre-whitened) so it is sufficient to search for an *orthogonal* matrix $\mathbf{W}$ (e.g., Hyvärinen et al., 2001). Additionally, jointly Gaussian sources are not identifiable under linear transformations, so we assume that no more than one source is normally distributed.

There are many approaches to solving the ICA problem, differing both in the objective function designed to measure the independence between the unmixed sources (sometimes referred to as a contrast function) and the optimization methods for that function. Most commonly used objective function is the mutual information (MI)

$$J(\mathbf{W}) = I\left(Y^1, \ldots, Y^d\right) = \sum_{i=1}^{d} H\left(Y^i\right) - H\left(Y^1, \ldots, Y^d\right) \tag{1}$$

where H is the Shannon entropy. Alternatively, one can minimize the sum $\sum_{i=1}^{d} H\left(\boldsymbol{y}^i\right)$ of the univariate entropies as the joint entropy is constant (e.g., Hyvärinen et al., 2001). Neither of these quantities can be evaluated directly, so approximations are used instead. Among effective methods falling in the former category is Kernel-ICA (Bach & Jordan, 2002); RADICAL (Learned-Miller & Fisher, 2003) and Fast-ICA (Hyvärinen, 1999) approximate the sum of the univariate entropies. There are other possible cost functions including maximum likelihood, moment-based methods, and correlation-based methods.

While ICA problems has been well-studied in the above formulation, there are a number of variations of it that are subject of active research. One such formulation is a noisy version of ICA

$$\mathbf{X} = \mathbf{AS} + \boldsymbol{\epsilon} \tag{2}$$

where multivariate noise $\boldsymbol{\epsilon}$ is often assumed normally distributed. Another related problem occurs when the mixed samples $\mathbf{X}$ are corrupted by a presence of outliers. There many other possibilities that go beyond the scope of this paper.

Of a special note is a generalization of ICA where some of the sources are *dependent*, independent subspace analysis (ISA). For this case, the mutual information and Shannon entropies from Equation 1 would involve multivariate random vectors instead of scalars. Resulting multidimensional entropies are exponentially more difficult to estimate than their scalar counterparts, making ISA problem more difficult than ICA. However, Cardoso (1998) conjectured that the ISA problem can be solved by first preprocessing the mixtures $\mathbf{X}$ by an ICA algorithm and then grouping the estimated components with highest dependence. The extent of this conjecture is still an open issue although it has been rigorously proven for some distribution types (Szabó et al., 2007). Although there is no proof for general sources as of yet, a number of algorithms apply this heuristics with success (Cardoso, 1998; Theis, 2007; Bach & Jordan, 2003).

3. Non-parametric Rank-Based Approach

Most of the ICA algorithms use an approximation to mutual information (MI) as their objective functions, and the quality of the solution thus depends on how accurate is the corresponding approximation. The problem with using MI is that without a parametric assumption on the functional form of the joint distribution, MI cannot be evaluated exactly, and numerical estimation can be both inaccurate and computationally expensive. In this section, we explore other measures of pairwise association as possible ICA contrasts. To note, most commonly used measure of correlation, Pearson's linear correlation coefficient cannot be used as it is invariant to rotations (once the data has been centered and whitened)

Instead, we are focusing on measures of dependence of the *ranks*. Ranks have a number of desirable properties – they are invariant under any monotonic transformations of the individual variables, insensitive to outliers, and not very sensitive to small amounts of

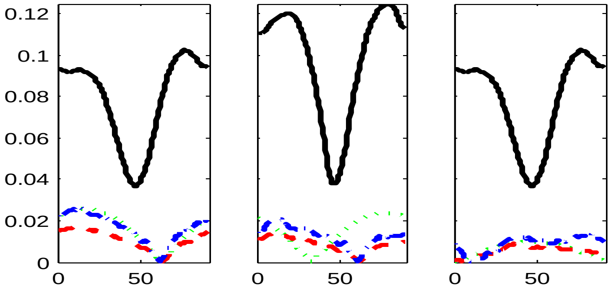


Figure 1. Values for sample version of Pearson's ρ_p (dotted, green), Kendall's τ (dashed, red), Spearman's ρ (dash-dotted, blue), and Schweizer-Wolff σ (solid, black) as a function of rotation angle $\left[0, \frac{\pi}{2}\right]$. Data was generated from a uniform distribution on $\mathbf{I}^2$ and rotated by $\frac{\pi}{4}$ (left), with added outliers (center), and with added noise (right).

noise. We found that a dependence measure defined on copulas (e.g., Nelsen, 2006), probability distributions on continuous ranks, has the right properties to be used as a contrast for ICA demixing.

3.1. Ranks and Copulas

Let a pair of random variables $(X, Y) \in \mathbb{R}^2$ be distributed according to a bivariate probability distribution P. Assume we are given N samples of (X, Y), $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$. Let the rank $r_x(x)$ be the number of x_i, $i = 1, \ldots, N$ such that $x > x_i$, and let $r_y(y)$ be defined similarly.

Many non-linear dependence measures are based on ranks. Among most commonly used are Kendall's τ and Spearman's ρ rank correlation coefficients. Kendall's τ measures the difference between proportions of concordant pairs ((x_i, y_i) and (x_j, y_j) such that $(x_i - x_j)(y_i - y_j) > 0$) and discordant pairs. Spearman's ρ measures a linear correlation between ranks of $r_x(x)$ and $r_y(y)$. Both τ and ρ have a range of $[-1, 1]$ and are equal to 0 (in the limit) if the X and Y are independent. However, the converse is not true, and both τ and ρ can be 0 even if X and Y are not independent. While they are robust to outliers, neither ρ nor τ make for a good ICA contrast as they provide a noisy estimate for dependence from moderately-sized data sets when the dependence is weak (See Figure 1 for an illustration).

Rank correlations can be extended from samples to distributions with the help of *copulas*, distributions over continuous multivariate ranks. Using a related to Spearman's ρ measure of dependence for copulas, we will devise an effective robust contrast for ICA.

Let $\mathbf{I}$ denote a unit interval $[0, 1]$. A bivariate *copula* C is probability function (cdf) defined on a unit square, $C : \mathbf{I}^2 \rightarrow \mathbf{I}$ such that its univariate marginals are uniform, i.e., $C(u, 1) = u$, $C(1, v) = v$, $\forall u, v, \in \mathbf{I}$.[2] Let $U = P_x(X)$ and $V = P_y(Y)$ denote the corresponding cdfs for previously defined random variables X and Y. Variables $X = P_x^{-1}(U)$ and $Y = P_y^{-1}(V)$ can be defined in terms of the inverse of marginal cdfs. Then, for $(u, v) \in \mathbf{I}^2$, define C as

$$C(u, v) = P\left(P_x^{-1}(u), P_y^{-1}(v)\right).$$

It is easy to verify that C is a copula. Sklar's theorem (Sklar, 1959) states that such copula exists for any distribution P, and that it is unique on the range of values of the marginal distributions. A copula can be thought of as binding univariate marginals P_x and P_y to make a distribution P.

Copulas can be viewed as a canonical form of multivariate distributions as they preserve multivariate dependence properties of the corresponding families of distributions. For example, the differential mutual information of the joint distribution is equal to the negentropy of its copula restricted to the region on which the copula *density* function (denoted in this paper by $c(u, v)$) is defined:

$$\begin{aligned} c(u, v) &= \frac{\partial^2 C(u, v)}{\partial u \partial v} = \frac{p(x, y)}{p_x(x)\, p_y(y)}; \\ I(X, Y) &= \int_{\mathbf{I}^2} c(u, v) \ln c(u, v)\, \mathrm{d}u \mathrm{d}v. \end{aligned}$$

Such negentropy is minimized when $C(u, v) = \Pi(u, v) = uv$. Copula Π is referred to as the *product* copula and is equivalent to variables U and V (and the original variables X and Y) being mutually independent. This copula will play a central part in definition of contrasts in the next subsection.

Copulas can also be viewed as a joint distribution over univariate *ranks*, and therefore, preserve all of the rank statistics of the corresponding multivariate distributions; rank based statistics can be expressed in terms of the copula alone. For example, Kendall's τ and Spearman's ρ have a convenient functional form in terms of the corresponding copulas (e.g., Nelsen, 2006):

$$\begin{aligned} \tau &= 4 \int_{\mathbf{I}^2} C(u, v)\, \mathrm{d}C(u, v) - 1, \\ \rho &= 12 \int_{\mathbf{I}^2} \left(C(u, v) - \Pi(u, v)\right) \mathrm{d}u \mathrm{d}v. \qquad (3) \end{aligned}$$

[2] While we restrict our attention to bivariate copulas, many of the definitions and properties described in this section can be extended to a d-variate case.

As the true distribution P and its copula C are not known, the rank statistics can be estimated from the available sample using an *empirical copula* (Deheuvels, 1979). For a data set $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, and empirical copula C_N is given by

$$C_N\left(\frac{i}{N}, \frac{j}{N}\right) = \frac{\#\text{ of } (x_k, y_k) \text{ s.t. } x_k \leq x_i \text{ and } y_k \leq y_j}{N}. \tag{4}$$

Well-known sample versions of several non-linear dependence measures can be obtained using an empirical copula (e.g., Nelsen, 2006). For example, sample version r of Spearman's ρ appears to be a grid integration evaluation of its expression in terms of a copula (Equation 3):

$$r = \frac{12}{N^2-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(C_N\left(\frac{i}{N}, \frac{j}{N}\right) - \frac{i}{N} \times \frac{j}{N}\right). \tag{5}$$

3.2. Schweizer-Wolff σ and λ

Part of the problem with Kendall's τ and Spearman's ρ as a contrast for ICA is a property that their value may be 0 even though the corresponding variables X and Y are not independent. Instead, we suggest using Schweizer-Wolff σ (Schweizer & Wolff, 1981), a measure of association from the class of *measures of dependence* (e.g., Nelsen, 2006, p. 208). This dependence measure can be viewed as an L_1 norm between a copula for the distribution and a product copula:

$$\sigma = 12\int_{\mathbf{I}^2} |C(u,v) - uv| \, \mathrm{d}u\mathrm{d}v. \tag{6}$$

σ has a range of $[0, 1]$, with an important property that $\sigma = 0$ if and only if the corresponding variables are mutually independent, i.e., $C = \Pi$. (This is one of seven requirements for measures of dependence.) The latter property suggests both an independence test and an ICA contrast for a pair of variables: pick a rotation angle such that the corresponding demixed data set has its Schweizer-Wolff (SW) dependence measure σ minimized. A sample version of σ is similar to a corresponding version of ρ (Equation 5):

$$s = \frac{12}{N^2-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|C_N\left(\frac{i}{N}, \frac{j}{N}\right) - \frac{i}{N} \times \frac{j}{N}\right|. \tag{7}$$

We note that other measures of dependence can be potentially used as an ICA contrast. We also experimented with an L_∞ version of σ, $\lambda = \sup_{\mathbf{I}^2} |C(u,v) - uv|$, a dependence measure similar to Kolmorogov-Smirnov univariate statistic, with results similar to SW σ.

4. SWICA: A New Algorithm for ICA and ISA

In this section, we present a new algorithm for ICA and ISA demixing. The algorithm uses Schweizer-Wolff (SW) σ estimates as a contrast in demixing pairs of variables; we named this algorithm Schweizer-Wolff contrast for ICA, or SWICA for short.

4.1. 2-dimensional Case

First, we tackle the case of a two-dimensional signal $\mathbf{S}$ mixed with a 2×2 matrix $\mathbf{A}$. We, further assume $\mathbf{A}$ is orthogonal (otherwise achievable by whitening). The problem is then reduced to finding a demixing rotation matrix $\mathbf{W} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$.

For the objective function, we use s (Equation 7) computed on $2 \times N$ matrix $\mathbf{Y} = \mathbf{WX}$ of rotated samples. Given an angle θ, $s(\boldsymbol{Y}(\theta))$ can be computed by first sorting each of the rows of $\mathbf{Y}(\theta)$ and computing row ranks for each entry of $\mathbf{Y}(\theta)$, then computing an empirical copula C_N (Equation 4) for ranks of $\mathbf{Y}$, and finally computing $s(\mathbf{Y}(\theta))$ (Equation 7). The solution is then found by finding angle θ minimizing $s(\mathbf{Y}(\theta))$. Similar to RADICAL (Learned-Miller & Fisher, 2003), we find such solution by searching over K values of θ in the interval $\left[0, \frac{\pi}{2}\right)$ (SW σ is invariant to rotations of the data by the angle measured in an integer multiples of $\frac{\pi}{2}$). This algorithm is outlined in Figure 2.

4.2. d-dimensional Case

A d-dimensional linear transformation described by a $d \times d$ orthogonal matrix $\mathbf{W}$ is equivalent to a composition of 2-dimensional rotations (called *Jacobi* or *Givens* rotations) (e.g., Comon, 1994). The transformation matrix itself can be written as a product of corresponding rotation matrices, $\mathbf{W} = \mathbf{W}_L \times \ldots \times \mathbf{W}_1$ where each matrix $\mathbf{W}_l$, $l = 1, \ldots, L$ is a rotation matrix (by angle θ_l) for some pair of dimensions (i, j). Thus a d-dimensional ICA problem can be solved by solving 2-dimensional ICA problems in succession. Given a current demixing matrix $\mathbf{W}_c = \mathbf{W}_l \times \ldots \times \mathbf{W}_1$ and a current version of the signal $\mathbf{X}_c = \mathbf{W}_c\mathbf{X}$, we find an angle θ corresponding to $\mathrm{SWICA}\left(\mathbf{X}_c^{(i,j)}, K\right)$. Taking an approach similar to RADICAL, we perform a fixed number of successive sweeps through all possible pairs of dimensions (i, j).

We should note that while d-dimensional SWICA is not guaranteed to converge, it converges in practice vast majority of the time. A likely explanation is that

```
Algorithm SWICA($\mathbf{X}, K$)
Inputs: $\mathbf{X}$, a $2 \times N$ matrix where rows are mixed signals (centered and whitened), $K$ equispaced evaluation angles in the $[0, \pi/2)$ interval

For each of $K$ angles $\theta$ in the interval $[0, \pi/2)$
($\theta = \frac{\pi k}{2}, k = 0, \ldots, K-1.$)
    • Compute rotation matrix
          $\mathbf{W}(\theta) = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$
    • Compute rotated signals $\mathbf{Y}(\theta) = \mathbf{W}(\theta)\mathbf{X}$.
    • Compute $s(\mathbf{Y}(\theta))$, a sample estimate of SW $\sigma$ (Equation 7)

Find best angle $\theta_m = \arg\min_\theta s(\mathbf{Y}(\theta))$

Output: Rotation matrix $\mathbf{W} = \mathbf{W}(\theta_m)$, demixed signal $\mathbf{Y} = \mathbf{Y}(\theta_m)$, and SW $\sigma$ estimate $s = s(\boldsymbol{Y}(\theta_m))$
```

Figure 2. Outline of SWICA algorithm (2-d case).

each 2-dimensional optimization finds a transformation that reduces the sum of entropies for the corresponding dimensions, reducing the overall sum of entropies. In addition to this, Learned-Miller and Fisher (2003) suggest that the minimization of the overall sum of entropies in this fashion (by changing only two terms in the sum) may make it easier to escape local minima.

4.3. Complexity Analysis and Acceleration Tricks

2-dimensional SWICA requires a search over K angles. For each angle, we first sort the data to compute the ranks of each data point ($\mathcal{O}(N \log N)$), and then use these ranks to compute $\hat{\delta}$ by computing the empirical copula and summing over the $N \times N$ grid (Equation 7), requiring $\mathcal{O}(N^2)$ additions. Therefore, running time complexity of 2-d SWICA is $\mathcal{O}(KN^2)$. Each sweep of a d-dimensional ICA problem solves a 2-dimensional ICA problem for each pair of variables, $\mathcal{O}(d^2)$ of them; S sweeps would have $\mathcal{O}(Sd^2KN^2)$ complexity. In our experiments, we employed $K = 180$ and $S = d - 1$.

The most expensive computation in SWICA is $\mathcal{O}(N^2)$ needed to compute SW estimate $s(\mathbf{Y}(\theta))$. Reducing this complexity, either by approximation, or perhaps, by an efficient rearrangement of the sum, is left to future research. We used several other tricks to speed up the computation. One, for large N ($N > 2500$) we estimated s by averaging s computed from bootstrapped samples of smaller size $N_b < N$. This approach reduces complexity to $\mathcal{O}(KBN_b^2)$ where B is the number of bootstrap samples. Two, when searching for θ minimizing $s(\mathbf{Y}(\theta))$, it is unnecessary to sum over all N^2 terms when evaluating a candidate θ if a partial sum already results in a value of $s(\mathbf{Y}(\theta))$ larger than the current best. This optimization translates to a 2-fold speed increase in practice. Three, it is unnecessary to complete all S sweeps if the algorithm already converged. One possible measure of convergence is an Amari error (Equation 8) measured for the cumulative rotation matrix for the most recent sweep.

4.4. Using Schweizer-Wolff σ for ISA

Schweizer-Wolff σ measure of dependence can be effectively used to convert ICA solutions into those for ISA by grouping together variables with high estimated values of SW σ. This grouping approach was proposed by Cardoso (1998), and while it is not provably correct, it is nonetheless effective. Our contribution is the use of the sample estimate of SW σ instead of the estimate of the mutual information.

5. Experiments

For the experimental evaluation of SWICA, we considered several settings. For the evaluation of the quality of demixing solution matrix $\mathbf{W}$, we computed the Amari error (Amari et al., 1996) for the resulting transformation matrix $\mathbf{B} = \mathbf{WA}$. Amari error $r(\mathbf{B})$ measures how different matrix $\mathbf{B}$ is from a permutation matrix, and is defined as

$$\alpha \sum_{i=1}^{d} \left(\frac{\sum_{j=1}^{d} |b_ij|}{\max_j |b_{ij}|} - 1 \right) + \alpha \sum_{j=1}^{d} \left(\frac{\sum_{i=1}^{d} |b_{ij}|}{\max_i |b_{ij}|} - 1 \right). \tag{8}$$

where $\alpha = 1/(2d(d-1))$. $r(\mathbf{B}) \in [0, 1]$, and $r(\mathbf{B}) = 0$ if and only if $\mathbf{B}$ is a permutation matrix. We compared SWICA to FastICA (Hyvärinen, 1999), Kernel-ICA (Bach & Jordan, 2002), RADICAL (Learned-Miller & Fisher, 2003), and JADE (Cardoso, 1999).

For the simulated data experiments, we used 18 different one-dimensional densities to simulate sources. These test-bed densities (and some of the experiments below) were proposed by Bach and Jordan (2002) to test Kernel-ICA and by Learned-Miller and Fisher (2003) to evaluate RADICAL; we omit the description of these densities due to lack of space as they can be looked up in the above papers.

Table 1 summarizes the medians of the Amari errors for 2-dimensional problems where both sources had the same distribution. Samples from these sources were then transformed by a random rotation, and then demixed using competing ICA algorithms. SWICA outperforms its competitors in 8 out of 20 cases, and performs comparably in several other cases. However, it performs poorly when the joint distribution for the sources is close to a Gaussian. One possible explanation for why SWICA performs worse than its competitors for these cases is that by using ranks instead of the actual values, SWICA is discarding some of the information that may be essential to separating such sources. However, given larger number of samples, SWICA is able to separate near-Gaussian sources (data not shown due to space constraints).

Figure 3 summarizes the performance of ICA algorithms in the presence of outliers for the 2-source case. Distributions for the sources were chosen at random from the 18 distributions from the experiment in Table 1. The sources were mixed using a random rotation matrix. The mixed sources were then corrupted by adding +5 or −5 to a single component for a small number of samples. SWICA significantly outperforms the rest of the algorithms as sample ranks used by SWICA are virtually unchanged by a small number of outliers. We tested SWICA further by significantly increasing the number of outliers; the performance was virtually unaffected when the proportion of the outliers was below 20%. SWICA is also less sensitive to noise than other ICA methods (Figure 4).

We further tested SWICA on sound and image data. We mixed $N = 1000$ samples from 8 sound pieces of an ICA benchmark [3] by a random orthogonal 8×8 matrix. Then we added 20 outliers to this mixture in the same way as in the previously described outlier experiment and demixed them using ICA algorithms. Figure 5 shows that SWICA outperforms other methods on this task. For the image experiment, we used 4 natural images[4] of size 128×256. The pixel intensities we normalized in the $[0,\ 255]$ interval. Each image was considered as a realization of a stochastic variable with 32768 sample points. We mixed these 4 images by a 4×4 random orthogonal mixing matrix, resulting in a mixture matrix of size 4×32768. Then we added large +2000 or −2000 outliers to 3% randomly selected points of these mixture, and then selected at random

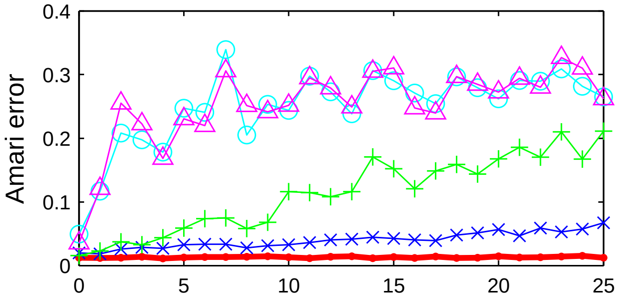


Figure 3. Amari errors for 2-dimensional ICA problem in the presence of outliers (multiplied by 100). The plot shows the median values over $K = 100$ repetitions of $N = 1000$ samples. SWICA is shown by red dots (thick), RADICAL by blue x, Kernel-ICA by green pluses, FastICA by cyan circles, and JADE by magenta triangles. The x-axis is the number of outliers (0 to 25).

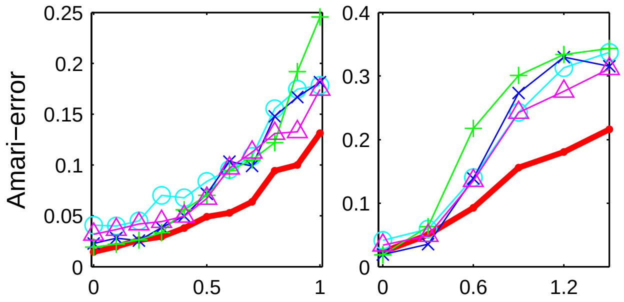


Figure 4. Amari errors (multiplied by 100) for 2-dimensional (left) and 6-dimensional ICA problems in the presence of independent Gaussian noise applied to mixed sources. The plot shows the median values of $K = 100$ repetitions of $N = 1000$ (left) and $N = 2000$ (right). The abscissa shows the variance σ^2 of the Gaussian noise, $\sigma^2 = (0, 0.1, \ldots, 0.9, 1)$ (left) and $\sigma^2 = (0, 0.3, 0.6, 0.9, 1.2, 1.5)$ (right). The legend is the same as in Figure 3.

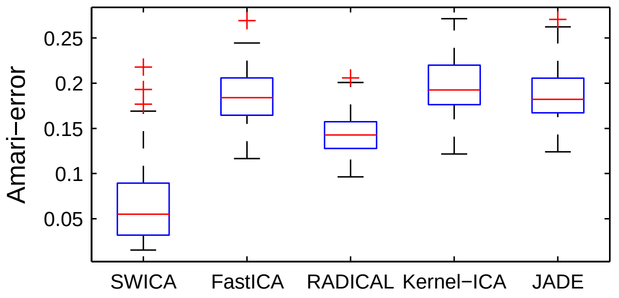


Figure 5. Box plot of Amari error for the mixed sounds with outliers. Plot computed over $K = 100$ runs.

[3] http://www.cis.hut.fi/projects/ica/cocktail/cocktail_en.cgi

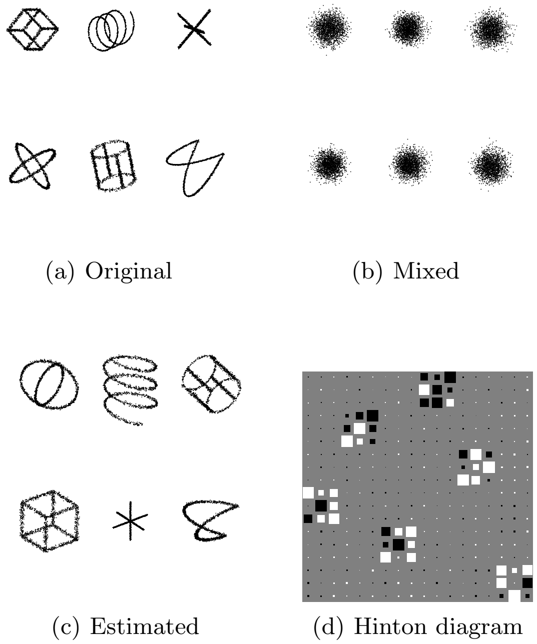
[4] http://www.cis.hut.fi/projects/ica/data/images/

Table 1. The Amari errors for two-component ICA with 1000 samples (multiplied by 100). Each entry is the median of 100 replicates for each pdf, (a) to (r). "rand" row represents the median of 1000 replicates when both densities were chosen uniformly at random from (a)-(r). The lowest (best) entry in each row is boldfaced.

pdf	SW	FastICA	Radical	KernelICA	Jade
a	3.74	3.01	2.18	**2.09**	2.67
b	2.39	4.87	**2.31**	2.50	3.47
c	**0.79**	1.91	1.60	1.54	1.63
d	10.10	5.63	4.10	5.05	**3.94**
e	**0.47**	4.75	1.43	1.21	3.27
f	**0.78**	2.85	1.39	1.34	2.77
g	**0.74**	1.49	1.19	1.11	1.19
h	3.66	5.32	4.01	3.54	**3.36**
i	10.21	7.38	6.95	7.70	**6.41**
j	**0.86**	4.64	1.29	1.21	3.38
k	**2.10**	5.58	2.65	2.38	3.53
l	4.09	7.68	**3.61**	3.65	5.21
m	**1.11**	3.41	1.43	1.23	2.58
n	2.08	4.05	2.10	**1.56**	4.07
o	5.07	3.81	2.86	2.92	**2.78**
p	**1.24**	2.92	1.81	1.53	2.70
q	3.01	12.84	2.30	**1.67**	10.78
r	3.32	4.30	3.06	**2.65**	3.32
rand	**1.47**	3.94	2.12	1.89	3.22

2000 samples from the 32768 vectors. We estimated the demixing matrix **W** using only these 2000 points, and then recovered the hidden sources for all 32768 samples using this matrix. SWICA significantly outperformed other methods. Figure 6 shows an example of the demixing achieved by different ICA algorithms.

Finally, we applied SW σ in an ISA setting. For this experiment, we used 6 3-dimensional sources where each of 3-dimensional variables were sampled from a geometric form (Figure 7a), resulting in a 18-dimensional hidden source. We then mixed this source ($N = 1000$ samples) with a random 18×18 orthogonal matrix (Figure 7b). We then applied Cardoso's conjecture. After processing the mixed sources using FastICA, the recovered sources were clustered with SW σ (Figure 7c). (We used a commonly used trick applying a non-linear transformation to recovered sources before grouping them. The dependence between independent sources is not affected by such transformations, but can be amplified for dependent sources.) We were able to recover the hidden subspaces with high precision as indicated by the Hinton diagram of **WA** (Figure 7d).


(a) Original

(b) Mixed

(c) Estimated

(d) Hinton diagram

Figure 7. ISA experiment for 6 3-dimensional sources.

6. Conclusion

We proposed a new ICA and ISA method, SWICA, based on a non-parametric rank-based estimate of the dependence between pairs of variables. The method frequently outperforms other state of the art ICA algorithms, is very robust to outliers, and only moderately sensitive to noise. On the other hand, it is somewhat slower than other ICA methods, and requires more samples to separate near-Gaussian sources. In the future, we plan to investigate possible accelerations to the algorithm, and statistical characteristics of the source distributions that affect the contrast.

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). Beyond independent components: Trees and clusters. *Journal of Machine Learning Research*, *4*, 1205–1233.

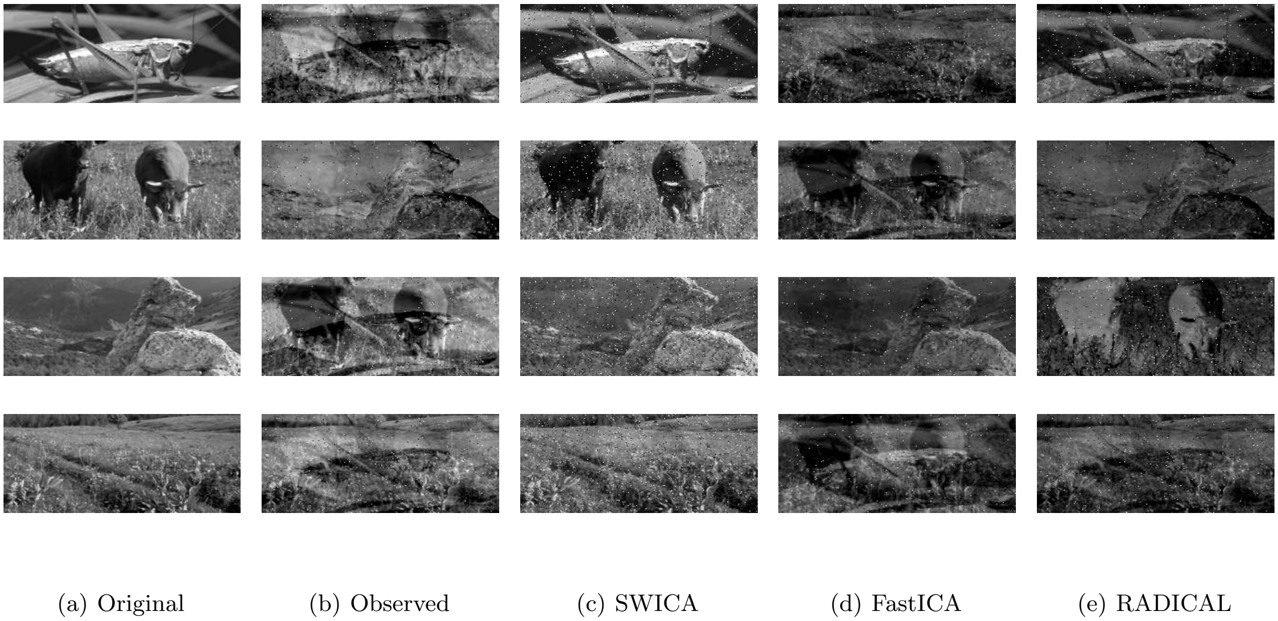


Figure 6. Separation of outlier-corrupted mixed images. (a) The original images. (b) the mixed images corrupted with outliers. (c)-(e) The separated images using SWICA, FastICA, and RADICAL algorithms, respectively. The Amari error of the SWICA, FastICA, Radical was 0.10, 0.30, 0.29 respectively. The quality of the Kernel-ICA and JADE was similar to that of FastICA and RADICAL.

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

Cardoso, J.-F. (1999). High-order contrasts for independent component analysis. *Neural Computation*, *11*, 157–192.

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

Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés, un test non paramétrique d'indépendance. *Bulletin de l'Académie Royale de Belgique, Classe des Sciences*, 274–292.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 626–634.

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

Hyvärinen, A., & Köster, U. (2006). FastISA: A fast fixed-point algorithm for independent subspace analysis. *Proc. of ESANN.* Evere, Belgium.

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

Nelsen, R. B. (2006). *An introduction to copulas.* Springer Series in Statistics. Springer. 2nd edition.

Póczos, B., & Lőrincz, A. (2005). Independent subspace analysis using geodesic spanning trees. *Proc. of ICML-2005* (pp. 673–680).

Schweizer, B., & Wolff, E. F. (1981). On nonparametric measures of dependence for random variables. *The Annals of Statistics*, *9*, 879–885.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leures marges. *Publications de l'Institut de Statistique de L'Université de Paris*, *8*, 229–231.

Szabó, Z., Póczos, B., & Lőrincz, A. (2007). Undercomplete blind subspace deconvolution. *Journal of Machine Learning Research*, *8*, 1063–1095.

Theis, F. J. (2005). Blind signal separation into groups of dependent signals using joint block diagonalization. *Proc. of ISCAS.* (pp. 5878–5881).

Theis, F. J. (2007). Towards a general independent subspace analysis. *Proc. of NIPS 19* (pp. 1361–1368).