EUSIPCO 05 - 13th European Signal Processing Conference
Antalya (Turkey), 4-8 September 2005, pp. 1412:1-4

A SIMPLE ICA ALGORITHM FOR NON-DIFFERENTIABLE CONTRASTS

John A. Lee, Frédéric Vrins and Michel Verleysen

Machine Learning Group, Université catholique de Louvain (UCL)
Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium
E-mail: {lee,vrins,verleysen}@dice.ucl.ac.be
Web: http://www.ucl.ac.be/mlg

ABSTRACT

A general-purpose algorithm is proposed for Independent Component Analysis. This algorithm is specifically designed in order to handle non-differentiable contrast functions. Sources are extracted one at a time (deflation approach). Examples are given for recently published contrast functions, e.g. the support width of the estimated source distribution.

1. INTRODUCTION

Independent Component Analysis (ICA) has been an active research field for more than a decade [3, 7]. Many ICA algorithms share a common model in which unknown source signals $\mathbf{s}$ are assumed to be statistically independent random variables. Observed signals $\mathbf{y}$ are then modeled as linear and instantaneous mixtures of the source signals:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \ , \tag{1}$$

where $\mathbf{A}$ is the *m*-by-*n* mixing matrix ($n < m$). Starting from this model, ICA aims at finding a separation matrix $\mathbf{B}$ such that a good approximation of the source signals can be found:

$$\mathbf{s} \approx \mathbf{y} = \mathbf{B}\mathbf{x} \ . \tag{2}$$

In this equation, the approximation symbol means not only that sources may be imperfectly recovered but also accounts for intrinsic indeterminacies of the model [3] (scaling and permutation).

Building an ICA algorithm from the above model requires firstly to formulate a contrast function $\mathscr{C}(\mathbf{y})$ that estimates the 'level of statistical independence' between the components of $\mathbf{y}$. Unfortunately, statistical independence is very difficult to measure and many ICA contrast functions actually computes quantities related to 'side effects' of statistical independence. For example, mutual information can be a contrast function for ICA [5]. Other contrast functions are related to the fact that mixtures of non-Gaussian sources signals are 'more Gaussian' than the sources. Therefore, measures of non-Gaussianity like negentropy [6], high-order cumulants [8], etc. can help to recover the independent sources in the ICA model.

Once a contrast function has been defined, the ICA algorithm has to maximize it. Usually, contrast functions are chosen in order to be differentiable because doing so allows using well known optimization tools like (natural) gradient ascent [2] or fixed-point techniques [8]. This practice leaves the field of non-differentiable contrast functions less explored, due to the difficulty of associating them with appropriate optimization techniques.

This paper proposes a simple and general-purpose algorithm for such contrast functions; it is organized as follows. Section 2 defines some preliminary concepts used in the algorithm. Next, Section 3 details the algorithm itself. Section 4 describes the experimental setting and gives the results of the algorithm with several contrast functions. Finally, Section 5 draws the conclusions.

2. FRAMEWORK AND DEFINITIONS

The ICA problem can be greatly simplified by preprocessing the observed variables $\mathbf{y}$ by Principal Component Analysis [9, 7] (PCA). PCA actually plays two roles: it allows both whitening the mixtures and reducing their number to the number of sources. As PCA yields uncorrelated mixtures $\mathbf{z}$ with zero mean and unit variance, matrix $\mathbf{B}$ can be factorized in the separation formula (2):

$$\mathbf{s} \approx \mathbf{y} = \mathbf{W}\mathbf{V}\mathbf{x} \ . \tag{3}$$

In this rewritten equation, $\mathbf{V}$ results from the PCA and is a *n*-by-*m* matrix defined as

$$\mathbf{V} = \mathrm{diag}([\lambda_1^{-1/2}, \ldots, \lambda_n^{-1/2}])\,[\mathbf{e}_1, \ldots, \mathbf{e}_n]^T \ , \tag{4}$$

where $\mathbf{e}_i$ is the eigenvector associated with the *i*-th largest eigenvalue λ_i of the covariance matrix $\mathbf{C}_{\mathbf{xx}}$. In 3, $\mathbf{W}$ is an orthogonal matrix ($\mathbf{W}^{-1} = \mathbf{W}^T$ and $\det \mathbf{W} = \pm 1$) and the only unknown part that remains in the ICA model. The algorithm proposed in this paper assumes that mixtures are preprocessed using PCA and is thus limited to the computation of the orthogonal matrix $\mathbf{W}$ starting from a set of $\mathbf{z}$ samples.

Contrast functions for ICA can be global (multi-unit) or componentwise (single-unit) [7]. In the first case, the function $\mathscr{C}(\mathbf{y})$ summarizes in one scalar value the level of independence between all pairs of components. In the second case, the function $\mathscr{C}(\mathbf{y}, i)$ measures a quantity related to the *i*-th component of $\mathbf{y}$, which is typically higher for independent signals than for mixtures of signals. For instance, mutual information belongs to the first category whereas negentropy and other non-Gaussianity measures like high-order cumulants belong to the second one. In the ICA jargon, the algorithms corresponding to global and componentwise contrasts are said to follow respectively a *symmetric* approach (all sources are recovered simultaneously) or a *deflation* approach (sources are extracted one after the other) [7]. The proposed algorithm follows the second approach and is thus only suited for componentwise contrasts. It works by performing small angular variations of the contrast argument. For this purpose, the componentwise contrast function $\mathscr{C}(\mathbf{y}, i)$ is rewritten as $\mathscr{C}(\mathbf{w}_i\mathbf{z})$, where $\mathbf{w}_i$ is the *i*-th row of $\mathbf{W}$. Knowing that $\mathbf{w}_i$ is orthogonal to any other row $\mathbf{w}_j$, positive and negative angular variations of $\mathbf{w}_i$ that preserve the unit norm may be noted and defined as

$$\mathbf{w}_{i\uparrow j} = \cos(\alpha)\,\mathbf{w}_i + \sin(\alpha)\,\mathbf{w}_j \ , \tag{5}$$

$$\mathbf{w}_{i\downarrow j} = \cos(\alpha)\,\mathbf{w}_i - \sin(\alpha)\,\mathbf{w}_j \ . \tag{6}$$

The corresponding contrast values can be written as $\mathscr{C}(\mathbf{w}_{i\uparrow j}\mathbf{z})$ and $\mathscr{C}(\mathbf{w}_{i\downarrow j}\mathbf{z})$.

In order to remain meaningful, the maximization procedure of the contrast function relies on the following assumptions. Firstly, the contrast function should be continuous or at least nearly continuous with respect to α: small variations of α should not cause disproportionate variations of the contrast. Next, all maxima of the contrast function should correspond to acceptable solutions of the ICA problem. On the other hand, it is not assumed that the contrast function is differentiable with respect to α. Therefore, the contrast function may be e.g. a piecewise linear function (discontinuous derivative), a staircase function (derivative in $\{-\infty, 0, +\infty\}$) or any other complicated function with no analytical derivative.

3. ALGORITHM

Under the above-mentioned assumptions, the simple algorithm shown in Fig. 1 may be used to compute each row of $\mathbf{W}$. Briefly put, for each row $\mathbf{w}_i$, the algorithm looks at the contrast value in each perpendicular direction ($\mathbf{w}_j$ with $i+1 < j < n$), for both positive and negative angular variations. Then it updates $\mathbf{W}$ by rotating both $\mathbf{w}_i$ and $\mathbf{w}_j$ according to the highest contrast value. As it can be seen, the algorithm keeps $\mathbf{W}$ orthogonal.

ICAFORNDC($\mathscr{C}, \mathbf{z}, \beta, \tau$)
Input: $\mathscr{C}$ (contrast function),
$\mathbf{z}$ (whitened mixtures),
β (convergence parameter, default: 0.75),
τ (number of iterations, default: 50).
Output: $\mathbf{W}$ (separation matrix).
Auxiliary: α (an angle),
n (number of sources), i, j, t (iteration indices).
Begin
▷ Initialize $\mathbf{W}$ to the identity matrix.
$\mathbf{W} \leftarrow \mathbf{I}_n$
▷ Deflation approach: estimate each source sequentially.
for $i \leftarrow 1$ **to** n **do**
  ▷ Iterate for the i-th source
  for $t \leftarrow 1$ **to** τ **do**
    ▷ Set the current angle variation.
    $\alpha \leftarrow \pi\beta^t$
    ▷ For each perpendicular direction.
    for $j \leftarrow i+1$ **to** n **do**
      ▷ Determine best contrast value.
      if $\mathscr{C}(\mathbf{w}_{i\uparrow j}\mathbf{z}) > \mathscr{C}(\mathbf{w}_i\mathbf{z})$ **and** $\mathscr{C}(\mathbf{w}_{i\uparrow j}\mathbf{z}) > \mathscr{C}(\mathbf{w}_{i\downarrow j}\mathbf{z})$ **then**
        ▷ Rotate the i-th and j-th rows of $\mathbf{W}$ accordingly ($+\alpha$).
        $\mathbf{w}_i, \mathbf{w}_j \leftarrow \mathbf{w}_{i\uparrow j}, \mathbf{w}_{j\downarrow i}$
      else if $\mathscr{C}(\mathbf{w}_{i\downarrow j}\mathbf{z}) > \mathscr{C}(\mathbf{w}_i\mathbf{z})$ **and** $\mathscr{C}(\mathbf{w}_{i\downarrow j}\mathbf{z}) > \mathscr{C}(\mathbf{w}_{i\uparrow j}\mathbf{z})$ **then**
        ▷ Rotate the i-th and j-th rows of $\mathbf{W}$ accordingly ($-\alpha$).
        $\mathbf{w}_i, \mathbf{w}_j \leftarrow \mathbf{w}_{i\downarrow j}, \mathbf{w}_{j\uparrow i}$
      end if
    end for
  end for
end for
Return $\mathbf{W}$
End

Figure 1: Pseudo-code for the deflation ICA algorithm for non-differentiable contrast functions (comments begin with a '▷').

The only parameters of the algorithm are β and τ. Usually, with the default values, the algorithm has converged after ten or twenty iterations ($10 < t < 20$). By construction, the algorithm is monotonic: the contrast is either increased or kept constant. For this reason, if spurious maxima of the contrast exist, then the algorithm could fall in one of them, especially if the initial values of α are too small or if α decreases too fast during the first iterations.

Similar algorithms using more sophisticated techniques, like discrete gradient approximations (based on a second-order Taylor expansion), have been tried too. Unfortunately, they lead to worse results than the simple proposed algorithm. In addition, they involve a larger number of parameters, which are tedious for adjusting.

4. EXPERIMENTS

Five ways to perform ICA are compared in this section. They combine three contrast functions and three optimization algorithms.

4.1 Contrast functions

4.1.1 Absolute kurtosis.

The (normalized) kurtosis (or kurtosis excess) of a random variable y with zero mean and unit variance can be written as

$$\kappa(y) = \frac{E\{y^4\}}{E\{y^4\}^2} - 3 = E\{y^4\} - 3 \quad . \tag{7}$$

and measures the non-Gaussianity of the random variable y (for a Gaussian variable $\kappa(y) = 0$). Therefore, the absolute kurtosis of an estimated source $\mathbf{w}_i\mathbf{z}$ may be used as an ICA contrast function [8, 7]. Obviouly, the kurtosis is differentiable and is used here for comparison purposes only, i.e. to show the differences between the proposed algorithm and another widely used deflation algorithm (FastICA).

4.1.2 Support width.

Previous work (Support Width Measure [11]) has shown the interest of the support width as a contrast for ICA. Actually, if sources are known to be bounded, minimizing the support width of the distribution of an estimated source solves the ICA problem. Knowing a set of obervations, the support width of a random variable y can be estimated by

$$\mathrm{SW}(y) = \max(y) - \min(y) \quad . \tag{8}$$

This approximation of the support width is very sensitive to noise and outliers. In order to make them more robust, minimum and maximum values may be replaced with the averages of the p smallest and largest values (quantiles may be used too [10]). In the experiments, p is equal to one percent of the number of observations. Minimum and maximum (and their robust approximations) are clearly non-differentiable functions.

4.1.3 Kullback-Leibler divergence.

Many componentwise contrasts are actually non-Gaussianity measures, like the absolute kurtosis. Knowing that the Kullback-Leibler divergence [4] (KL) measures a pseudo-distance (it is not symmetric) between two probability density functions (PDFs), it can be used as an ICA contrast if the KL divergence is measured between a Gaussian distribution and a an estimated source. Of course, the KL divergence is difficult to estimate because in its continuous form, it requires to know the source PDFs and to integrate them. Fortunately, the KL divergence between two random variables y_1 and y_2 may be approximated using normalized histograms and the discrete formula:

$$\mathrm{KL}(y_1, y_2) = \sum_k^B b_k(y_1) \log\left(\frac{b_k(y_1)}{b_k(y_2)}\right) \quad , \tag{9}$$

where B is the number of bins in the histogram and b_k is the normalized height of the k-th bin ($\sum_{k=1}^B b_k(y_i) = 1$). In the experiments, the histogram includes 32 bins in the interval $[-6, +6]$.

4.2 Algorithms

Three ICA algorithms are used: (i) the above-described procedure for non-differentiable contrasts, (ii) FastICA, version 2.3, deflation approach, fine tuning enabled, (iii) do nothing after whitening, i.e. $\mathbf{W}$ equal to identity. FastICA is used with the absolute kurtosis ('pow3' nonlinearity), which is the only differentiable contrast listed in the previous section. The proposed algorithm is used to

- Maximize the absolute kurtosis (AKMICA, $\mathscr{C}(\mathbf{w}_i\mathbf{z}) = |\kappa(\mathbf{w}_i\mathbf{z})|$).
- Minimize the support width (SWICA, $\mathscr{C}(\mathbf{w}_i\mathbf{z}) = -\mathrm{SW}(\mathbf{w}_i\mathbf{z})$).
- Maximize the discrete approximation of the Kullback-Leibler divergence (KLICA, $\mathscr{C}(\mathbf{w}_i\mathbf{z}) = \mathrm{KL}(\mathbf{w}_i\mathbf{z}, g)$ where $g \sim N(0,1)$).

4.3 Mixtures

Five mixtures are artificially generated from five sources. These sources are:

1. A sine wave: $\sin(13\pi t/1000)$.
2. A triangular signal (sawteeth): $\arcsin(\sin(17\pi t/1000))$.
3. Observations drawn from a χ^2 distribution with three degrees of freedom.
4. Observations drawn from a Student's t distribution with five degrees of freedom.
5. Observations drawn from a normal distribution with zero mean and unit variance.

Thousand observations are computed ($t \in \{1,2,3,\ldots,1000\}$ for the sine wave and the sawteeth) or drawn at random (for the three other distributions). Next, the observations are standardized, as in Fig. 2: the mean of each source is subtracted and each source is divided by its standard deviation. Then these observations are mixed using a 5-by-5 matrix whose entries are chosen at random in a normal distribution with zero mean and unit variance. Finally, the mixtures are whitened as decribed in Section 2 before running the five abovementioned ICA algorithms.

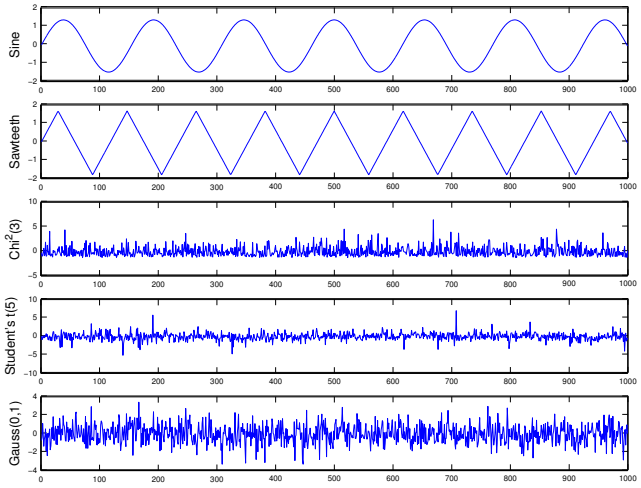

Figure 2: Artificially generated sources: a sine wave (bimodal distribution), a triangular sawteeth signal (uniform distribution), observations drawn from a $\chi^2(3)$ distribution, from a Student's $t(5)$ distribution and from a Gaussian $N(0,1)$ normal distribution.

4.4 Performance assessment

The sources estimated by the five ICA algorithms are compared to the original sources using the source-to-interference ratio (SIR), inspired from [1]. For this purpose, the equation system $\mathbf{y} = \mathbf{C}\mathbf{s}$ is solved and for each source the SIR is computed as

$$\mathrm{SIR}(y_i) = \frac{\sum_j |c_{ij}|}{\max_i |c_{ij}|} - 1 \quad , \tag{10}$$

where $y_i = \mathbf{w}_i\mathbf{z}$ and c_{ij} is the entry of $\mathbf{C}$ at the crossing of the i-th row and j-th column. The SIRs are computed for each source separately and then summed to obtain a global performance indicator.

4.5 Results

Five hundreds different mixtures have been generated and separated. The histograms of the SIR values are shown in Fig. 3 for each source and for all sources. Mean and standard deviations of the SIR for the 500 trials are given in Table 1. The first conclusion to draw is that FastICA remains the fastest ICA algorithm and offers an appealing trade-off between speed, quality (low mean of summed SIRs) and robustness (low standard deviation of summed SIRs). The AKMICA algorithm reaches almost as good results and even equals FastICA regarding the robustness. This clearly shows that (i) the kurtosis remains an excellent contrast function for ICA algorithms, (ii) fixed-point (and probably other algorithms using analytic gradients) are superior to simpler procedures like the proposed one. Nevertheless, the latter offers other advantages. Firstly, it may be used with a wide variety of contrast functions. Secondly, these contrast functions can yield excellent results for specific types of sources. SWICA clearly outperforms all other studied algorithms for abruptly bounded sources like the sine wave or the sawteeth (uniform distribution). In a similar way, KLICA seems particularly suited for assymetric sources like the χ^2 distribution or bimodal distributions like the sine wave. All in all, KLICA even performs better than FastICA on the proposed mixtures (lowest average of summed SIRs) but lacks some robustness: the algorithm sometimes fails to converge properly (high standard deviation of the summed SIRs). This good performance, however, requires a sufficient number of observations and a careful adjustment of the number of bins in the histogram involved in the discrete approximation of the Kullback-Leibler divergence.

5. CONCLUSION

A simple ICA algorithm, based on a deflation approach of the ICA problem and suited for non-differentiable componentwise contrast function has been studied. Experiments have shown that even a very simple optimization procedure can solve an ICA problem and yield high quality results. Because the proposed algorithm is not limited to differentiable contrasts (like gradient-based algorithms), new possibilities may be tried. For instance, the support width of the distribution of the estimated source succeeds very well in recovering abruptly bounded sources. Similarly, a histogram-based estimate of the Kullback-Leibler divergence between the distribution of the estimated source and a standardized Gaussian distribution leads to good results too, in particular for assymmetric or bimodal sources.

Future works aims at improving the performance and robustness of the proposed algorithm and at developing a symmetric version.

REFERENCES

[1] S. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind signal separation. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 757–763. MIT Press, Cambridge, MA, 1995.

[2] J.-F. Cardoso and B. H. Laheld. Equivariant adaptative source separation. *IEEE Transactions on Signal Processing*, 44(12):3017–3030, 1996.

[3] P. Comon. Independent Component Analysis – A new concept? *Signal Processing*, 36:287–314, 1994.

[4] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley, New York, NY, 1991.

[5] A. Hyvärinen. Independent component analysis by minimization of mutual information. Technical Report A46, Helsinki University of Technology, Laboratory of Computer and Information Science, 1997.

[6] A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. Technical Report A47, Helsinki University of Technology, Laboratory of Computer and Information Science, 1997.

[7] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley-Interscience, 2001.

[8] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9:1483–1492, 1997.

[9] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, NY, 1986.

[10] D.-T. Pham. Blind separation of instantenaous mixtrures of sources based on order statistics. *IEEE Transactions on Signal Processing*, 48(2):363–375, 2000.

[11] F. Vrins, C. Jutten, and M. Verleysen. SWM: A class of convex contrasts for source separation. In *Proceedings of ICASSP'05, IEEE International Conference on Acoustics, Speech, and Signal Processing*. Philadelphia, PA, March 2005. Accepted.

EUSIPCO 05 - 13th European Signal Processing Conference
Antalya (Turkey), 4-8 September 2005, pp. 1412:1-4

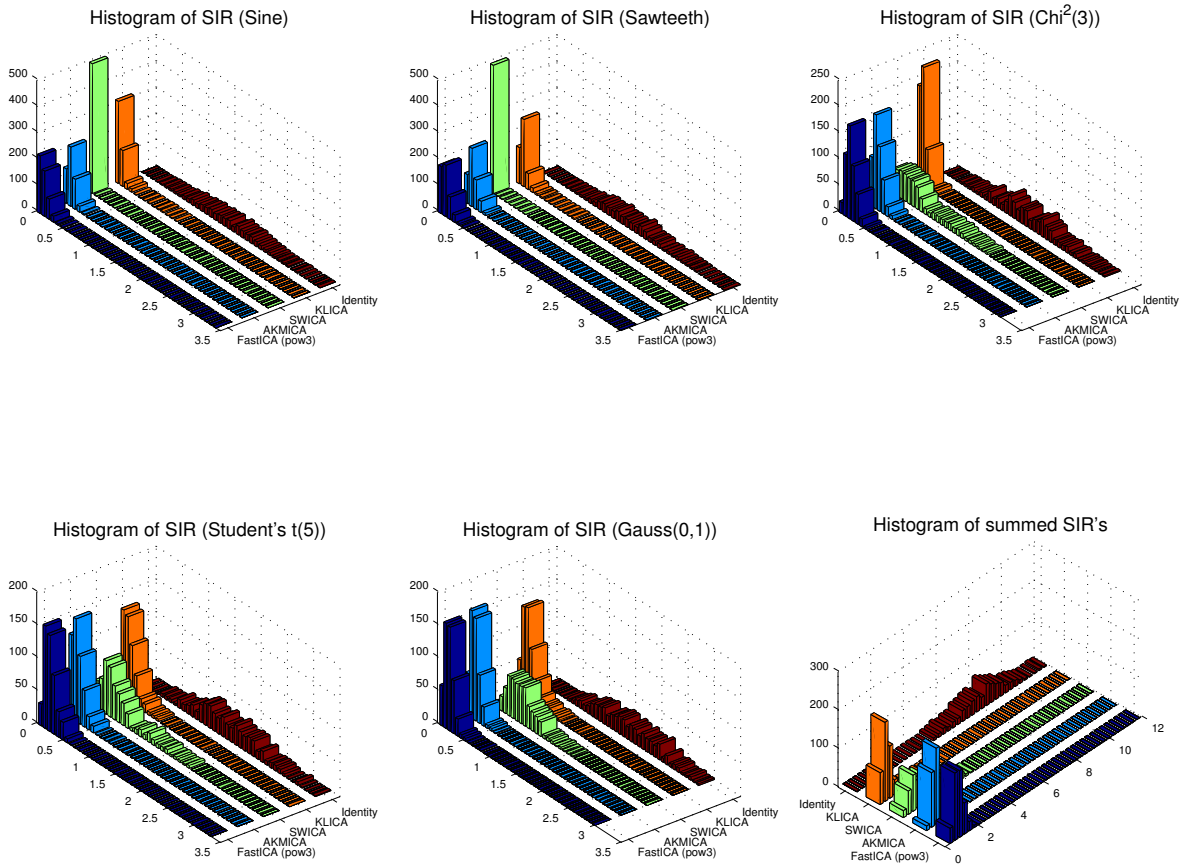

Figure 3: Histograms of SIR values for 500 ICA trials. Source-to-interference ratio (SIR) is shown for each source and for all sources (sum of SIRs).

	Sine	Sawteeth	$\chi^2(3)$	Student's $t(5)$	Normal $N(0,1)$	All sources
FastICA	0.1145 (0.0835)	0.1354 (0.0833)	0.2264 (0.0937)	**0.2404** (**0.1097**)	0.2042 (0.0919)	0.9208 (0.2943)
AKMICA	0.1359 (0.0699)	0.1585 (0.0802)	0.2450 (0.1022)	0.2588 (0.1134)	**0.2013** (**0.0876**)	0.9995 (**0.2919**)
SWICA	**0.0060** (**0.0390**)	**0.0302** (**0.0503**)	0.5276 (0.4305)	0.5402 (0.3782)	0.5787 (0.2460)	1.6827 (0.9390)
KLICA	0.0962 (0.1496)	0.1596 (0.1621)	**0.1173** (**0.0557**)	0.2556 (0.1525)	0.2351 (0.1327)	**0.8638** (0.4438)
Identity	1.6030 (0.5532)	1.6228 (0.5427)	1.5820 (0.5486)	1.6019 (0.5425)	1.6068 (0.5485)	8.0164 (1.4207)

Table 1: Mean and standard deviation of the SIR values for 500 ICA trials. Values are given for each source and for all sources (sum of SIRs).