Proceedings of the Fourth International Symposium on Independent Component Analysis and Blind Signal Separation, 2003

ICA USING SPACINGS ESTIMATES OF ENTROPY

Erik G. Miller[*]

University of California at Berkeley
Dept. of Elect. Eng. and Comp. Sci.
Berkeley, CA 94720, USA

John W. Fisher III[†]

Massachusetts Institute of Technology
Dept. of Elect. Eng. and Comp. Sci.
Cambridge, MA 02139, USA

ABSTRACT

This paper presents a new algorithm for the independent components analysis (ICA) problem based on efficient *spacings estimates* of entropy. Like many previous methods, we minimize a standard measure of the departure from independence, the estimated Kullback-Leibler divergence between a joint distribution and the product of its marginals. To do this, we use a consistent and rapidly converging entropy estimator due to Vasicek. The resulting algorithm is simple, computationally efficient, intuitively appealing, and outperforms other well known algorithms. In addition, the estimator and the resulting algorithm exhibit excellent robustness to outliers. We present favorable comparisons to Kernel ICA, FAST-ICA, JADE, and extended Infomax in extensive simulations.

1. INTRODUCTION

We present a new independent components analysis (ICA) algorithm. Empirical results indicate that it outperforms a wide array of well known algorithms. Several principles its development:

1. Since ICA is, by definition, about maximizing statistical independence, we attempt to directly optimize a measure of statistical independence, rather than a surrogate for this measure.

2. We avoid explicit estimation of probability densities as an intermediate step. Indeed, given the formulation of the objective function, density estimation (even implicitly) is entirely unnecessary.

3. Since our objective function involves one-dimensional entropy estimation, we employ a consistent, rapidly converging and computationally efficient estimator of entropy which is robust to outliers. For this task, we turned to the statistics literature, where entropy estimators have been studied extensively (c.f. [1]).

4. As the optimization landscape has potentially many local minima, we eschew gradient descent methods. The computational efficiency of our estimator allows for a global search method. The properties of the ICA problem allow extension of this technique to higher dimensions in a tractable manner.

Attention to these principles led to the Robust, Accurate, Direct ICA aLgorithm (RADICAL) presented here.

ICA as applied to instantaneous linear mixtures considers the generative model of random observations [2]

$$X = AS. \tag{1}$$

Here $X \in \Re^C$ and $S \in \Re^D$ are random vectors, and $A \in \Re^{C\times D}$ is a fixed but unknown mixing matrix. We will assume that the mixing matrix A has full rank, the components of S are mutually independent, and $C = D$, or A is square. The goal is to recover (in some sense) the sources and perhaps the mixing matrix via a transformation W on observations of X, that is

$$Y = WX = WAS = BS. \tag{2}$$

In this setting, it has been shown [3, 2] that one can recover the original sources up to a scaling and permutation provided that at most one of the underlying sources is Gaussian and the rest are non-Gaussian. See [4] for an extensive bibliography of the ICA problem.

Many approaches start the analysis of the problem by considering the contrast function [2]

$$\begin{align} & J(Y) \tag{3} \\ = & \int p(y_1,\cdots,y_D)\log\frac{p(y_1,\cdots,y_D)}{p(y_1)p(y_2)...p(y_D)}\;d\mu \tag{4} \\ = & \sum_{i=1}^{D} H(Y_i) - H(Y_1,\cdots,Y_D), \tag{5} \end{align}$$

where $d\mu = dy_1 dy_2 \cdots dy_D$ and $H(Y)$ is the differential entropy of the continuous multidimensional random variable Y. Since (4) will be 0 if and only if all of the variables

[*] Supported by ONR grant N00014-01-1-0890 under the MURI program. The authors would like to thank Francis Bach for substantial help in comparing RADICAL to other algorithms.

[†] Supported by Army Research Office under grant DAAD19-00-1-0466 and Office of Naval Research under grant N00014-00-1-0907.

are *mutually independent*, we take (4) as a direct measure of mutual independence.

As a function of X and W it is easily shown (c.f. [5]) that

$$J(Y) = \sum_{i=1}^{D} H(Y_i) - H(X_1, \ldots, X_D) - \log(|W|),$$

i.e., the change in the joint entropy under linear transformation is simply the logarithm of the Jacobian of the transformation. We will assume the X's are pre-whitened, and hence W will be restricted to rotation matrices (i.e. $\log(|W|) = 0$) and the minimization of $J(Y)$ reduces to finding

$$W^* = \arg\min_{W} H(Y_1) + \cdots + H(Y_D). \qquad (6)$$

To estimate this quantity, we adopt a different entropy estimator, almost identical to one described by Vasicek [6] and others[1] in the statistics literature. These estimators, which are discussed below, are known as *spacings estimates*.

2. SPACINGS ESTIMATES OF ENTROPY

Consider a one-dimensional random variable Z, and a random sample of Z denoted by $Z^1, Z^2, ..., Z^N$. The *order statistics* of a random sample of Z are simply the elements of the sample rearranged in non-decreasing order: $Z^{(1)} \le Z^{(2)} \le ... \le Z^{(N)}$. A *spacing of order m*, or m*-spacing*, is then defined to be $Z^{(i+m)} - Z^{(i)}$, for $1 \le i < i+m \le N$. Finally, if m is a function of N, one may define the m_N*-spacing* as $Z^{(i+m_N)} - Z^{(i)}$.

The m_N-spacing estimator of entropy, originally due to [6], can now be defined as

$$\hat{H}_N(Z^1, ..., Z^N) = \frac{1}{N} \sum_{i=1}^{N-m_N} \log\left(\frac{N}{m_N}(Z^{(i+m_N)} - Z^{(i)})\right). \qquad (7)$$

This estimator is nearly equivalent to the one used in RADICAL. To see where it comes from, we make the following observation regarding order statistics. For *any random variable Z with impulse-free density $p(\cdot)$ and continuous distribution $P(\cdot)$*, the following holds. Let p^* be the N-way product density $p^*(Z^1, ..., Z^N) = p(Z^1)p(Z^2)...p(Z^N)$. Then

$$E_{p^*}[P(Z^{(i+1)}) - P(Z^{(i)})] = \frac{1}{N+1}, \quad \forall i, 1 \le i \le N-1. \qquad (8)$$

That is, the expected value of the probability mass of the interval between two successive elements of a sample from a random variable[2] is just $\frac{1}{N+1}$. This surprisingly general fact is a simple consequence of the uniformity of the random variable $P(Z)$, the random variable describing the "height" on the cumulative curve of a random draw from Z. $P(Z)$ is called the *probability integral transform* of Z. The key insight is that the *intervals* between successive order statistics tend to have about the same amount of probability mass.

Using this idea, one can develop a simple entropy estimator. We start by approximating the probability density $p(z)$ by assigning equivalent masses to each interval between points and assuming a uniform distribution of this mass across the interval[3]. Defining $Z^{(0)}$ to be the infimum of the support of $p(z)$ and defining $Z^{(N+1)}$ to be the supremum of the support of $p(z)$, we have:

$$\hat{p}(z; Z^1, ..., Z^N) = \frac{\frac{1}{N+1}}{Z^{(i+1)} - Z^{(i)}}, \qquad (9)$$

for $Z^{(i)} \le z < Z^{(i+1)}$. Then, we can write

$$\begin{aligned}
& H(Z) \\
= \quad & -\int_{-\infty}^{\infty} p(z) \log p(z) dz \\
\overset{(a)}{\approx} \quad & -\int_{-\infty}^{\infty} \hat{p}(z) \log \hat{p}(z) dz \\
= \quad & -\sum_{i=0}^{N} \int_{Z^{(i)}}^{Z^{(i+1)}} \hat{p}(z) \log \hat{p}(z) dz \\
= \quad & -\sum_{i=0}^{N} \int_{Z^{(i)}}^{Z^{(i+1)}} \frac{\frac{1}{N+1}}{Z^{(i+1)} - Z^{(i)}} \log \frac{\frac{1}{N+1}}{Z^{(i+1)} - Z^{(i)}} dz \\
= \quad & -\sum_{i=0}^{N} \frac{\frac{1}{N+1}}{Z^{(i+1)} - Z^{(i)}} \log \frac{\frac{1}{N+1}}{Z^{(i+1)} - Z^{(i)}} \int_{Z^{(i)}}^{Z^{(i+1)}} dz \\
= \quad & -\frac{1}{N+1} \sum_{i=0}^{N} \log \frac{\frac{1}{N+1}}{Z^{(i+1)} - Z^{(i)}} \\
\overset{(b)}{\approx} \quad & -\frac{1}{N-1} \sum_{i=1}^{N-1} \log \frac{\frac{1}{N+1}}{Z^{(i+1)} - Z^{(i)}} \\
= \quad & \frac{1}{N-1} \sum_{i=1}^{N-1} \log\left((N+1)(Z^{(i+1)} - Z^{(i)})\right) \\
\equiv \quad & \hat{H}_{simple}(Z^1, ..., Z^N).
\end{aligned}$$

The approximation (a) arises by approximating the true density $p(z)$ by $\hat{p}(z; Z^1, ..., Z^N)$. The approximation (b) stems from the fact that we in general do not know $Z^{(0)}$ and $Z^{(N+1)}$, i.e. the true support of the unknown density. The estimate (10) has both intuitive and theoretical appeal[4], but it has relatively high variance since while the expectation of

[1] For a review of these estimators and other entropy estimators, see [1].

[2] The probability mass of the interval between two successive points is equivalent to the integral of the density function between these two points.

[3] We use the notion of a density estimate to aid in the intuition behinid $m-$spacing estimates of entropy. However, as we stress below, density estimation *is not* a necessary intermediate step in entropy estimation.

[4] The addition of a small constant renders this estimator weakly consistent for bounded densities under certain tail conditions ([7]).

the interval probabilities (8) is $\frac{1}{N+1}$, their variance is high. This problem can be mitigated, and asymptotically eliminated completely, by considering $m-$spacing estimates of entropy, such as

$$\hat{H}_{m-spacing}(Z^1, ..., Z^N) \equiv \qquad (10)$$
$$\frac{m}{N-1} \sum_{i=0}^{\frac{N-1}{m}-1} \log\left(\frac{N+1}{m}(Z^{(m(i+1)+1)} - Z^{(mi+1)})\right).$$

By letting

$$m \rightarrow \infty, \frac{m}{N} \rightarrow 0, \qquad (11)$$

this estimator also becomes consistent ([1]). In this work, we typically set $m = \sqrt{N}$.

The intuition behind this estimator is that by considering m-spacings with larger and larger values of m, the variance of the probability mass of these spacings, relative to their expected value, gets smaller and smaller.

A modification of (10) in which the $m-$spacings overlap:

$$\hat{H}_{RADICAL}(Z^1, ..., Z^N) \equiv \qquad (12)$$
$$\frac{1}{N-m} \sum_{i=1}^{N-m} \log\left(\frac{N+1}{m}(Z^{(i+m)} - Z^{(i)})\right),$$

is used in RADICAL. This is asymptotically equivalent to the estimator (7) of [6]. Weak and strong consistency have been shown given (11) by various authors under a variety of general conditions. For details of these results, see the review paper [1]. Perhaps the most important property of this estimator is that it is asymptotically efficient, as shown in [8].

We note that Pham [9] defined an ICA contrast function as a sum of terms very similar to (10). However, by choosing $m = 1$ as was done in that work, one no longer obtains a consistent estimator of entropy, and the efficiency and efficacy of (10) as an ICA contrast function appears to be greatly reduced. In particular, much of the information about whether or not the marginals are independent is ignored in such an approach.

3. RADICAL IN TWO DIMENSIONS

Given that we have an entropy estimate (12) in hand, we now discuss its application to optimizing (6), starting with the two-dimensional ICA problem. Two issues which must be dealt with are local minima and “false minima”. Local minima are intrinsic to the optimization criterion (6) and persist even in the case of infinite data, i.e. when the entropy estimates are exact. False minima, on the other hand, are minima in (6) due to poor estimates of the entropy based on small sample sizes. We start by addressing false minima.

False minima can become quite severe when the sample size is small and are a consequence of the fact that the m-spacings estimator makes no prior smoothness assumptions (e.g. limited spatial frequency) regarding the underlying densities. Consequently, for small sample sizes there exist rotations (instances of W) for which portions of the data approximately align, producing an artificial spike in one of the marginal distributions. This phenomenon is easily understood by considering the case in which m, the number of intervals combined in an m-spacing, is equal to 1. In this case, for any value of N there are many rotations $W(\theta)$ which will cause two points to align exactly, either vertically or horizontally. This causes the $1-$spacing corresponding to these two points to have width 0, which in turn gives this interval an average logarithm of $-\infty$. This results in an entropy estimate of $-\infty$ for this particular rotation of the data. The entropy estimator has no evidence that there is not, in fact, an impulse in the true marginal density which would legitimately indicate a negatively infinite entropy, so it is not a fundamental flaw with the estimator. Rather, it is a consequence of allowing arbitrarily peaked implicit marginal estimates. While this issue becomes less of a problem as N and m grow, our empirical findings suggest that for the densities considered in this paper, it must be addressed to achieve good performance at least while $N \leq 1000$.

Consequently, we consider a smoothed version of the estimator. We attempt to remove such false minima by synthesizing R replicates of each of the original N sample points with additive spherical Gaussian noise to make a surrogate data set X'. That is, each point X^j is replaced with R samples from the distribution $N(X^j, \sigma^2 I)$, where R and σ^2 become parameters of the algorithm. Then we use the entropy estimator (12) on the expanded data set X'.

Even if this initial smoothing effectively eliminates many false minima, we must still address the issue of true local minima of the cost function. Local minima arise, for example, when one or more of the original source distributions S_i are multimodal. For 2-D source separation we take advantage of the fact that $W(\theta)$ is a one-dimensional manifold to do an exhaustive search over W for K values of θ. Note that we need only consider θ in the interval $[0, \frac{\pi}{2}]$, since any 90 degree rotation will result in equivalent independent components. In our two-dimensional experiments, we set $K = 150$. Importantly, it turns out that even in higher dimensions, our algorithm will remain linear in K, so it is relatively inexpensive to do a finer grain search over θ if desired. Complexity issues will be discussed in more detail below.

In two dimensions, RADICAL is a very simple algorithm, which is summarized in Figure 1. The algorithm has four parameters. The first parameter m, determines the number of intervals combined in an $m-$spacing. As stated above, we chose $m = \sqrt{N}$ for all of our experiments, which

```
Input:      Data vectors $X^1, ..., X^N$, assumed whitened.
Params.:    $m$: Size of spacing. Usually equal to $\sqrt{N}$.
            $R$: Number of replicated points per
                original data point.
            $\sigma$: Standard deviation for replicated points.
            $K$: Number of angles at which to evaluate
                cost function.
Procedure:  1. Create $X'$ by replicating $R$ points with
               Gaussian noise for each original point.
            2. For each $\theta$, rotate the data to this angle
               and evaluate cost function.
            3. Output $W$ corresponding to optimal $\theta$.
Output:     W, the demixing matrix.
```

Fig. 1. RADICAL in two dimensions.

guarantees the asymptotic consistency of our procedure, as long as the original source densities are impulse free.

A second parameter is the number of points R used to replace each original point X^j when creating the augmented data set. We used a value of $R = 30$ for all of our two-dimensional experiments. We note again, however, that because the entropy estimator is consistent, for large N, R can be reduced to 1 (and the replication procedure eliminated). The rate at which R can be reduced as a function of N is dependent upon the densities of the specific components S. The standard deviation of the R added points for each of the N points X^j is given by σ. Performance was relatively robust to the choice of this parameter and we chose only two different values of σ for all of our experiments. For $N < 1000$, we set $\sigma = 0.35$ and for $N >= 1000$, we set $\sigma = 0.175$.

The only remaining parameter for RADICAL in two dimensions is K, the number of rotations at which to measure the objective function. Since the error metric is approximately proportional to the difference in angle between the estimated θ and the true "unmixing" θ, it is easy to see that asymptotically, a lower bound on the minimum expected error is approximately $\frac{\pi}{4K}$. However, this bound does not become relevant until N is large enough to give very accurate entropy estimates.

In informal experiments, we tried values for K of 50, 100, 150, and 250. There was no noticable improvement in performance for $K > 150$, even for $N = 4000$ and the higher dimensional tests. For $K = 150$, the lower bound on error is approximately 0.005. Since the errors for the experiments given here were much larger than this, there was no advantage in further increasing K. In other words, K need be no larger than is warranted by the size N of the data set. Note, however, that since both the two-dimensional and higher-dimensional versions of RADICAL are linear in K, it is relatively inexpensive to increase the resolution of the exhaustive search.

3.1. Algorithmic complexity

The complexity of RADICAL in two dimensions is a function of two main elements. First, each evaluation of the entropy estimator requires a sort of $|X| = N$ data points, or when point replication is used $|X'| = RN \equiv N'$ data points. This gives a complexity of $O(N \log N)$, or $O(N' \log N')$ for each evaluation of the cost function. When N is large enough, point replication becomes unnecessary, so asymptotically each evaluation is $O(N \log N)$.

Secondly, we evaluate the cost function K times. This results in an apparent final complexity of $O(KN \log N)$ for RADICAL in two dimensions. We note, however, that large savings can be obtained since resorting the N points after a slight rotation $d\theta$ can be done more efficiently than sorting the N points for the first time, since most points will be in the correct relative positions. Under certain conditions, this resortng can be done in $O(N)$ time which gives a total complexity of $O(N \log N + KN)$. However, for the most general scenario, we have yet to prove a complexity better than $O(KN \log N)$.

3.2. Experiments in two dimensions

To test the algorithm experimentally, we performed a large set of experiments, largely drawn from the comprehensive tests developed by Bach and Jordan [10]. Our tests included comparisons with FastICA [11], the JADE algorithm [12], the extended Infomax algorithm [13], and KernelICA using the generalized variance [10].

For 18 different one-dimensional densities,[5] the following experiments were performed. Using a density $q(\cdot)$, N points were drawn iid from the product distribution $q(\cdot)q(\cdot)$. The points were then subjected to a random rotation matrix A to produce the input X for the algorithm[6]. We then measured the "difference" between the true matrix A and the W returned by the algorithm, according to the well-known criterion (Amari error), due to Amari et al. [15].

Table 1 shows the mean results for each source density on each row, with $N = 250$, the number of input points, and 100 replications of each experiment. The best performing algorithm on each row is shown in bold face. Note that RADICAL performs best in 10 of 18 experiments, substantially outperforming the second best in many cases. The mean performance in these experiments is shown in the row labeled **mean**, where RADICAL has lower error than all other algorithms tested. The final row of the table represents experiments in which two (generally different) source densities were chosen randomly from the set of 18 densities

[5] These densities and additional details of the experiments are described in [14].

[6] Alternatively, we could have applied a random non-singular matrix to the data, and then whitened the data, keeping track of the whitening matrix. For the size of N in this experiment, these two methods are essentially equivalent.

pdfs	F-ica	Jade	Imax	KGV	RADICAL
a	8.9	7.5	56.3	5.7	**5.6**
b	10.2	9.3	61.8	**6.2**	7.0
c	4.4	3.1	18.4	4.3	**2.4**
d	11.8	**10.0**	61.1	11.6	12.6
e	8.1	7.4	67.7	3.1	**1.7**
f	7.9	5.5	12.4	3.3	**2.0**
g	3.9	2.9	18.1	2.9	**1.4**
h	11.1	**8.2**	27.2	8.4	12.1
i	18.5	**16.7**	37.6	23.2	27.0
j	12.2	12.8	50.5	3.0	**1.7**
k	14.1	10.3	30.2	**5.2**	5.5
l	22.6	16.4	39.2	**8.7**	11.7
m	8.2	6.9	29.5	12.3	**1.9**
n	11.4	9.7	32.1	9.7	**3.9**
o	8.7	**6.8**	23.7	9.4	8.6
p	9.9	6.7	29.1	6.0	**2.6**
q	35.8	32.0	39.1	9.4	**5.3**
r	13.0	9.5	27.7	**7.2**	8.9
mean	12.3	10.1	36.8	7.8	**6.8**
rand	10.7	8.5	29.6	6.0	**5.8**

Table 1. The Amari errors (multiplied by 100) for two-component ICA with 250 samples. For each density (see [10]), averages over 100 replicates are presented. For each distribution, the lowest error rate is shown in bold face. The overall mean is calculated in the row labeled **mean**. The **rand** row presents the average over 1000 replications when two (generally different) pdfs were chosen uniformly at random among the 18 possible pdfs.

to produce the product distribution from which points were sampled. 1000 replications were performed using these randomly chosen distributions. For these experiments, RADICAL has a slight edge over Kernel-ICA, but they both significantly outperform the other methods.

Figure 2 shows results for our outlier experiments. These experiments were again replications of the experiments performed by [10]. It can be seen that RADICAL is uniformly more robust to outliers than all other methods in these experiments, for every number of outliers added.

4. RADICAL IN D DIMENSIONS

We now discuss the extension of RADICAL to problems with dimension D greater than two. To find the $D-$dimensional rotation matrix W^* that optimizes (6) in D dimensions, we use Jacobi methods such as those used to solve symmetric eigenvalue problems, and as applied to the ICA problem in [2]. The basic idea is to rotate the augmented data X' two dimensions at a time using *Jacobi rotations* (c.f. [16]). Since a Jacobi rotation $J(p, q, \theta)$ in the (p, q) plane leaves all components of an D-dimensional data point X^j unchanged except for the pth and qth components, optimizing our objective function (6) reduces to a two-dimensional ICA problem for each distinct Jacobi rotation.

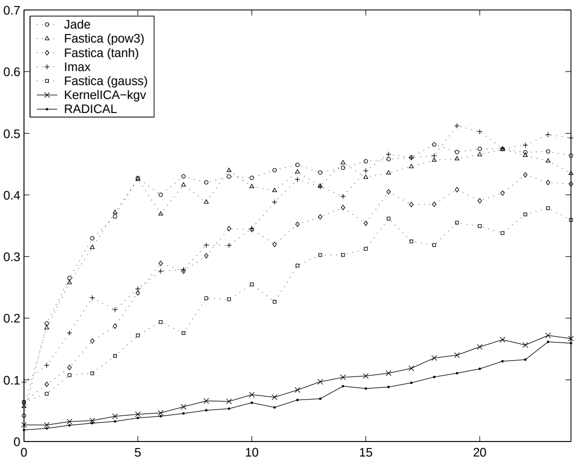

Fig. 2. Robustness to outliers. The abcissa displays the number of outliers and the ordinate shows the Amari error.

Algorithmically, we initialize Y to our augmented data set X', and our rotation matrix W to the identity matrix. After optimizing our objective function for a pair of dimensions (p, q), we update Y:

$$Y_{new} = J(p, q, \theta^*)Y, \tag{13}$$

keeping track of our cumulative rotation matrix:

$$W_{new} = J(p, q, \theta^*)W. \tag{14}$$

Note that since each Jacobi rotation affects only two components of Y, this is an $O(2^2 N') = O(N')$ operation. Thus, full scale $D-$dimensional rotations need never be done. There are $D(D-1)/2$ distinct Jacobi rotations (parameterized by θ), and performing a set of these is known as a *sweep*.

Empirically, performing multiple sweeps improves our estimate of W^* for some number of iterations, and after this point, the error may increase or decrease sporadically near its smallest value. The number of sweeps S becomes an additional parameter for multi-dimensional RADICAL. We found that with in dimension as high as sixteen, there was no additional improvement after $S = 8$ sweeps in our experiments. In practice, rather than setting a fixed value of this parameter, we iterated sweeps until the change in W was below some small tolerance.

To evaluate the complexity of RADICAL in D dimensions, we first note that there are $O(D^2)$ Jacobi rotations, rather than simply 1 rotation as there was in two dimensions. Second, the algorithm is linear in the number of sweeps S. Hence, the final complexity is at worst $O(SD^2 KN \log N)$.

dim	N	#repl	Fast	Jade	Imax	Kgv	Rad
2	250	1000	11	9	30	**5**	6
	1000	1000	5	4	7	**2**	**2**
4	1000	100	18	13	25	11	**6**
	4000	100	8	7	11	4	**3**
8	2000	50	26	22	123	20	**11**
	4000	50	18	16	41	**8**	**8**
16	4000	25	42	38	130	19	**15**

Table 2. Amari errors for experiments in higher dimension. The table shows experiments for dimensions two through 16. The number of points used for each experiment is shown in the second column and the number of replications performed to obtain the mean values at right is given in the third column. Note that RADICAL performed best or second best in every experiment, performing better than all other algorithms in four of seven experiments.

We note again that there are substantial opportunities for savings in computational cost over the most naive implementation. These will be discussed in future work.

It is also interesting to note that while Jacobi rotations have been applied in previous work, by combining them with a one-dimensional exhaustive search technique, we escape many of the local minima in which a traditional Jacobi algorithm would be trapped. For this benefit we pay only the price of replacing a gradient method of an unknown number of steps by the constant factor K.

Table 2 presents results of experiments for multiple dimensions. In each experiment for dimension D, D (generally) different densities were selected at random from the set of 18 densities discussed above. Samples from the resulting product distributions were again randomly rotated, and the task was to recover the independent components. Hence, our empirical results show that RADICAL exhibits excellent performance in two dimensions, in higher-dimensional problems, and also has excellent robustness to outliers.

5. REFERENCES

[1] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. van der Meulen, “Nonparametric entropy estimation: An overview,” *International Journal of Math. Stat. Sci.*, vol. 6, no. 1, pp. 17–39, June 1997.

[2] Pierre Comon, “Independent component analysis - a new concept?,” *Signal Processing*, vol. 36, pp. 287–314, 1994.

[3] A. Benveniste, M. Goursat, and G. Ruget, “Robust identification of a nonminimum phase system: blind adjustment of a linear equalizer in data communications,” *IEEE Transactions on Automatic Control*, vol. 25, no. 3, pp. 385–99, 1980.

[4] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, New York: John Wiley & Sons, 2001.

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

[6] Oldrich Vasicek, “A test for normality based on sample entropy,” *Journal of the Royal Statistical Society, Series B*, vol. 38, no. 1, pp. 54–59, 1976.

[7] P. Hall, “Limit theorems for sums of general functions of m-spacings,” *Math. Proc. Camb. Phil. Soc.*, vol. 96, pp. 517–532, 1984.

[8] B. Ya Levit, “Asymptotically optimal estimation of nonlinear functionals,” *Problems of Information Transmission*, vol. 14, pp. 65–72, 1978.

[9] D.-T. Pham, “Blind separation of instantaneous mixture of sources based on order statistics,” *IEEE Transactions on Signal Processing*, vol. 48, no. 2, pp. 363–375, 2000.

[10] Francis R. Bach and Michael I. Jordan, “Kernel independent component analysis,” *Journal of Machine Learning Research*, vol. 3, pp. 1–48, 2002.

[11] A. Hyvärinen and E. Oja, “A fast fixed point algorithm for independent component analysis,” *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[12] J.-F. Cardoso, “High-order contrasts for independent component analysis,” *Neural Computation*, vol. 11, no. 1, pp. 157–192, 1999.

[13] T.-W. Lee, M. Girolami, and T. J. Sejnowski, “Independent component analysis using an extended Infomax algorithm for mixed sub-gaussian and super-gaussian sources,” *Neural Computation*, vol. 11, no. 2, pp. 417–441, 1999.

[14] Erik G. Miller and John W. Fisher III, “Independent components analysis by direct entropy minimization,” Tech. Rep. UCB/CSD-03-1221, University of California at Berkeley, January 2003.

[15] S. Amari, A. Cichocki, and H. H. Yang, “A new learning algorithm for blind signal separation,” in *Advances in Neural Information Processing Systems, 8*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds. 1996, Cambridge, MA: MIT Press.

[16] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Baltimore, MD: Johns Hopkins University Press, 1996.