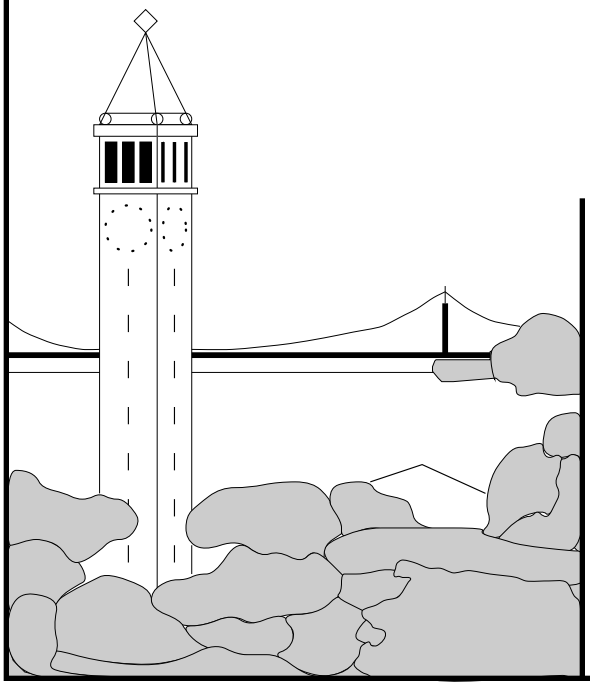
Independent Components Analysis by Direct Entropy Minimization

Erik G. Miller
egmil@eecs.berkeley.edu
Computer Science Division
University of California
Berkeley, CA 94720, USA

John W. Fisher III
fisher@ai.mit.edu
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
200 Technology Square, Office NE43-V 626
Cambridge MA 02139, USA



Report No. UCB/CSD-3-1221

January 2003

Computer Science Division (EECS)
University of California
Berkeley, California 94720

Independent Components Analysis by Direct Entropy Minimization

Erik G. Miller EGMIL@EECS.BERKELEY.EDU
Department of Electrical Engineering and Computer Science
University of California
Berkeley, CA 94720-1776, USA

John W. Fisher III FISHER@AI.MIT.EDU
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
200 Technology Square, Office NE43-V 626
Cambridge MA 02139, USA

Abstract

This paper presents a new algorithm for the independent components analysis (ICA) problem based on efficient entropy estimates. Like many previous methods, this algorithm directly minimizes the measure of departure from independence according to the estimated Kullback-Leibler divergence between the joint distribution and the product of the marginal distributions. We pair this approach with efficient entropy estimators from the statistics literature. In particular, the entropy estimator we use is consistent and exhibits rapid convergence. The algorithm based on this estimator is simple, computationally efficient, intuitively appealing, and outperforms other well known algorithms. In addition, the estimator's relative insensitivity to outliers translates into superior performance by our ICA algorithm on outlier tests. We present favorable comparisons to the Kernel ICA, FAST-ICA, JADE, and extended Infomax algorithms in extensive simulations.

1. Introduction

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

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 well-known[1], consistent, rapidly converging and computationally efficient estimator of

1. Although the estimator we use (Vasicek (1976)) has been extensively analyzed in the statistics literature, it appears to be relatively unknown in the machine learning community.

entropy which is robust to outliers. For this task, we turned to the statistics literature, where entropy estimators have been studied extensively (c.f. Beirlant et al. (1997)).

4. As the optimization landscape has potentially many local minima, we eschew gradient descent methods. The fact that the estimator is computationally efficient 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.

The paper is organized as follows. We begin by setting up the problem and discussing aspects of the contrast function originally proposed by Comon (1994) which can be simplified to a sum of one-dimensional marginal entropies (c.f. Kullback (1959)). In Section 2, we discuss entropy estimates based on order statistics. One method in particular satisfies our requirements, the *m-spacing* estimator (Vasicek, 1976).

This entropy estimator leads naturally to a simple ICA algorithm, a two-dimensional version of which is described in Section 3. We follow this with a discussion of complexity and experimental results for the two-dimensional problem. In addition to working for a wide variety of possible source distributions, we demonstrate that RADICAL has excellent robustness to the presence of outliers. In Section 4, we extend the algorithm to higher dimensional problems, with experiments in up to 16 dimensions. We discuss additional details of the algorithm and discuss why the complexity is still manageable, even using our exhaustive search approach.

1.1 Linear ICA and KL divergence

As articulated by Comon (1994), independent components analysis (ICA) or alternatively blind source separation as applied to instantaneous linear mixtures considers the generative model of random observations

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

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 (hence the term *blind*) mixing matrix. Typically, at a minimum, one assumes that

1. the mixing matrix A has full rank,
2. the components of S are mutually independent, and
3. $C \geq D$.

Condition 2 is equivalent to the statement that the joint density of the components of S can be expressed as a product of the marginals:

$$p(S_1, \cdots, S_D) = \prod_{i=1}^{D} p(S_i). \tag{2}$$

Additionally, here we shall restrict ourselves to the case where $C = D$ (i.e. A is square) without loss of generality. 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

$$\begin{aligned} Y &= WX && (3)\\ &= WAS && (4)\\ &= BS. && (5) \end{aligned}$$

Given the minimal statement of the problem, it has been shown (Benveniste et al., 1980, Comon, 1994) 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. Upon pre-whitening the observed data the problem reduces to a search over rotation matrices in order to recover the sources and mixing matrix in the sense described above (Hyvärinen, 2001, Bach and Jordan, 2002). We will assume henceforth that such pre-processing has been done.

While the problem statement is fairly straightforward with a minimum of assumptions it has been well studied, resulting in a vast array of approaches. Some of the more notable approaches can be roughly grouped into maximum likelihood based methods (Pham et al., 1992, Pearlmutter and Parra, 1996), moment/cumulant based methods (Comon, 1994, Cardoso and Souloumiac, 1996, Cardoso, 1999b, Hyvärinen, 2001), entropy based methods (Bell and Sejnowski, 1995, Hyvärinen, 1999), and correlation based methods (Jutten and Herault, 1991, Bach and Jordan, 2002).

Many approaches start the analysis of the problem by considering the contrast function (Comon, 1994)

$$\begin{aligned} J(Y) &= \int p(y_1, \cdots, y_D) \log \frac{p(y_1, \cdots, y_D)}{p(y_1)p(y_2)...p(y_D)}\, d\mu && (6)\\ &= KL\left(p(y_1, \cdots, y_D) || \prod_{i=1}^{D} p(y_i)\right) && (7)\\ &= \sum_{i=1}^{D} H(Y_i) - H(Y_1, \cdots, Y_D). && (8) \end{aligned}$$

Here $d\mu = dy_1 dy_2 \cdots dy_D$ and $H(Y)$ is the differential entropy (Shannon, 1948) of the continuous multi-dimensional random variable Y. The right side of (6) is the Kullback-Leibler divergence (Kullback, 1959), or relative entropy, between the joint density of $\{Y_1, \ldots, Y_D\}$ and the product of its marginals.

The utility of (6) for purposes of the ICA problem has been well documented in the literature (c.f. Comon (1994), Lee et al. (1999a)). Briefly we note that for mutually independent random variables $Y_1, Y_2, ..., Y_D$ we have:

$$\begin{aligned} J(Y) &= \int p(y_1, y_2, ..., y_D) \log \frac{p(y_1, y_2, ..., y_D)}{p(y_1)p(y_2)...p(y_D)} d\mu && (9)\\ &= \int p(y_1, y_2, ..., y_D) \log 1\, d\mu && (10)\\ &= 0. && (11) \end{aligned}$$

Since this quantity will be 0 if and only if all of the variables are *mutually independent*, we take (6) as a direct measure of mutual independence.

As a function of X and W it is easily shown (c.f. (Cover and Thomas, 1991, Bell and Sejnowski, 1995, Hyvärinen, 2001)) that

$$J(Y) \quad = \quad \sum_{i=1}^{D} H(Y_i) - H(X_1, \ldots, X_D) - \log\left(|W|\right). \tag{12}$$

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

$$W^* = \arg\min_{W} H(Y_1) + \cdots + H(Y_D). \tag{13}$$

The preceding development was necessary to bring us to the primary contribution of this paper. The observations noted in the development are the collective contributions of the cited authors. As has been noted (Hyvärinen, 1999), ICA algorithms consist of an objective (contrast) function *and* an optimization algorithm. We adopt the previously proposed objective criterion of (13) and present a means of both estimating its value and optimizing the choice of W via a method which is reliable, robust, and computationally efficient. These are the aspects of our proposed approach which will be the subject of the rest of the paper.

Towards that end, we adopt a different entropy estimator to minimize (13). The entropy estimator is almost identical to one described by Vasicek (1976) and others (c.f. Beirlant et al. (1997) for a review) in the statistics literature. This class of entropy estimators has not heretofore been applied to the ICA problem. As we will show, the use of this entropy estimator has a significant impact on performance as compared to other ICA algorithms and as discussed in the sections on experimental results. In addition, it has the following desirable properties:

- It is consistent.
- It converges as the square root of N, the number of data points, and is asymptotically efficient.
- It is computable in $O(N \log N)$ time.

In the next section, we present a detailed discussion of the entropy estimator and its properties.

2. Entropy Estimators for Continuous Random Variables

There are a variety of ICA algorithms that minimize (13) to find the independent components (e.g. Comon (1994)). These algorithms differ mostly in how they estimate the entropy of the one-dimensional marginal variables. Hyvärinen (1997), for example, developed a new entropy estimator for this purpose. RADICAL also uses entropy minimization at its core, and as such must estimate the entropy of each marginal for each possible W matrix. RADICAL's marginal entropy estimates are functions of the *order statistics* of those marginals.

2.1 Order statistics and spacings

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: $Z^{(1)} \leq Z^{(2)} \leq ... \leq Z^{(N)}$ (c.f. Arnold et al. (1992)). A *spacing of order m*, or *m-spacing*, is then defined to be $Z^{(i+m)} - Z^{(i)}$, for $1 \leq i < i + m \leq N$. 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 Vasicek (1976), can now be defined as

$$\hat{H}_N(Z^1, Z^2, ..., Z^N) = \frac{1}{N} \sum_{i=1}^{N-m_N} \log\left(\frac{N}{m_N}(Z^{(i+m_N)} - Z^{(i)})\right). \tag{14}$$

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

$$E_{p^*}[P(Z^{(i+1)}) - P(Z^{(i)})] = \frac{1}{N+1}, \quad \forall i, 1 \leq i \leq N-1. \tag{15}$$

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 $\frac{1}{N+1}$ of the total probability (which by definition must be equal to 1.0). This surprisingly general fact is a simple consequence of the uniformity of the random variable $P(Z)$. $P(Z)$, the random variable describing the "height" on the cumulative curve of a random draw from Z (as opposed to the function $P(z)$) is called the *probability integral transform* of Z (c.f. Manoukian (1986)). Thus, the key insight is that the *intervals* between successive order statistics have the same expected 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:

$$\hat{p}(z; Z^1, ..., Z^N) = \frac{\frac{1}{N+1}}{Z^{(i+1)} - Z^{(i)}}, \tag{16}$$

2. The probability mass of the interval between two successive points can also be thought of as 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.

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

$$
\begin{aligned}
H(Z) &= -\int_{-\infty}^{\infty} p(z) \log p(z) dz && (17)\\
&\overset{(a)}{\approx} -\int_{-\infty}^{\infty} \hat{p}(z) \log \hat{p}(z) dz && (18)\\
&= -\sum_{i=0}^{N} \int_{Z^{(i)}}^{Z^{(i+1)}} \hat{p}(z) \log \hat{p}(z) dz && (19)\\
&= -\sum_{i=0}^{N} \int_{Z^{(i)}}^{Z^{(i+1)}} \frac{\frac{1}{N+1}}{Z^{(i+1)} - Z^{(i)}} \log \frac{\frac{1}{N+1}}{Z^{(i+1)} - Z^{(i)}} dz && (20)\\
&= -\sum_{i=0}^{N} \frac{\frac{1}{N+1}}{Z^{(i+1)} - Z^{(i)}} \log \frac{\frac{1}{N+1}}{Z^{(i+1)} - Z^{(i)}} \int_{Z^{(i)}}^{Z^{(i+1)}} dz && (21)\\
&= -\frac{1}{N+1} \sum_{i=0}^{N} \log \frac{\frac{1}{N+1}}{Z^{(i+1)} - Z^{(i)}} && (22)\\
&\overset{(b)}{\approx} -\frac{1}{N-1} \sum_{i=1}^{N-1} \log \frac{\frac{1}{N+1}}{Z^{(i+1)} - Z^{(i)}} && (23)\\
&= \frac{1}{N-1} \sum_{i=1}^{N-1} \log \Big((N+1)(Z^{(i+1)} - Z^{(i)}) \Big) && (24)\\
&\equiv \hat{H}_{simple}(Z^1, ..., Z^N). && (25)
\end{aligned}
$$

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 in general we do not know $Z^{(0)}$ and $Z^{(N+1)}$, i.e. the true support of the unknown density. Therefore, we form the mean log density estimate using only information from the region for which we have some information, ignoring the intervals outside the range of the sample. This is equivalent to assuming that outside the sample range, the true density has the same mean log probability density as the rest of the distribution.

2.2 Lowering the variance of the estimate

The estimate (25) has both intuitive and theoretical appeal[4], but it has relatively high variance since while the expectation of the interval probabilities (15) is $\frac{1}{N+1}$, their variance is high. The upper left plot of Figure 1 shows the distribution of values obtained when we divide a random $1-$spacing by its expected value. Clearly, these values are widely distributed around the ideal value of 1.

4. The addition of a small constant renders this estimator weakly consistent for bounded densities under certain tail conditions (Hall (1984)).

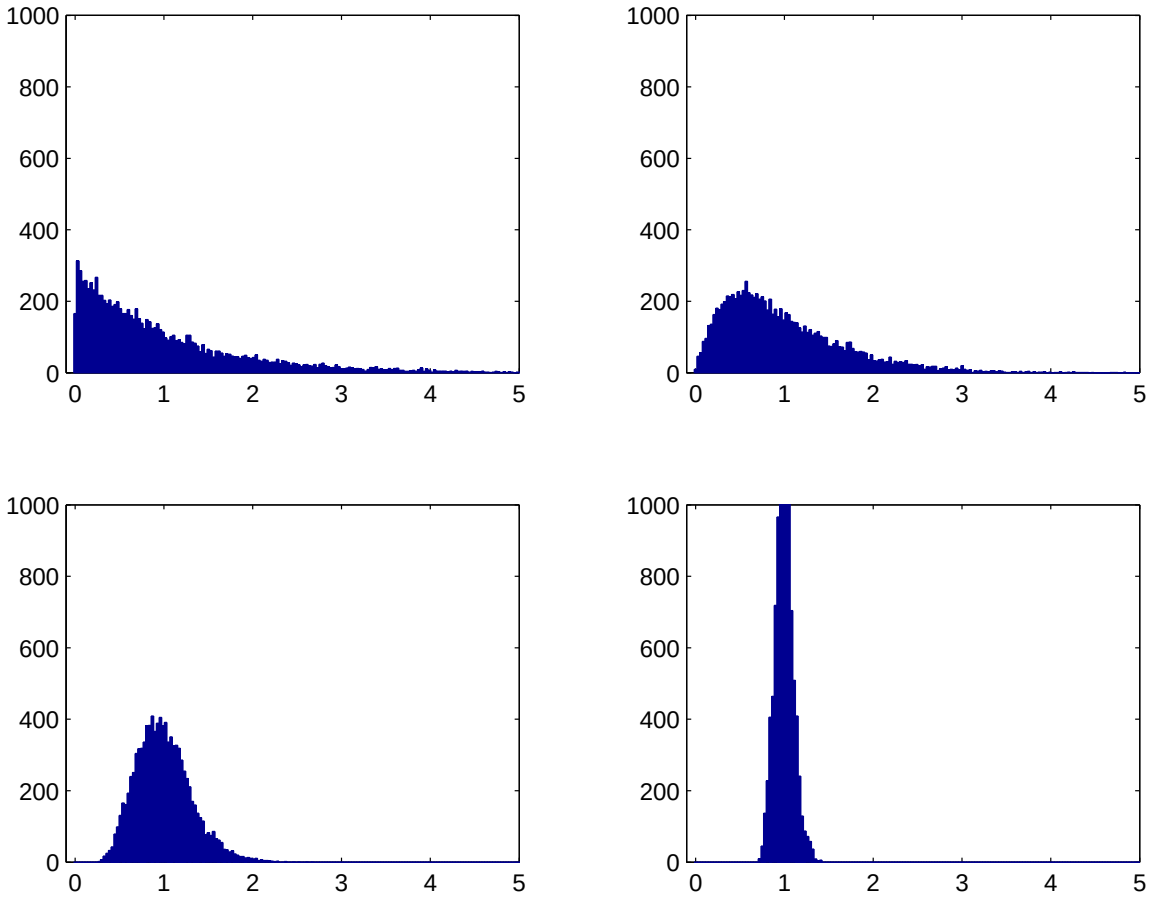


Figure 1: Histograms showing the variability of the probability mass of m-spacings, as a function of m. Each plot shows, for a particular m, the ratio of a set of random m-spacings to their expected values. When $m = 1$ (upper left plot), the probability mass of the m-spacings is widely distributed. Already for $m = 2$ (upper right), the ratio is substantially more concentrated around its expected value of 1. For $m = 10$ (lower left), the m-spacings' probability masses are almost always within a factor of three of their expected values. For $m = 100$ (lower right), the probability mass of the m-spacings is highly consistent. This behavior is a simple consequence of the law of large numbers and the uniformity of the distribution obtained through the probability integral transform.

This problem can be mitigated, and asymptotically eliminated completely, by considering $m-$spacing estimates of entropy, such as

$$\hat{H}_{m-spacing}(Z^1,...,Z^N) \equiv \frac{m}{N-1} \sum_{i=0}^{\frac{N-1}{m}-1} \log\left(\frac{N+1}{m}(Z^{(m(i+1)+1)} - Z^{(mi+1)})\right). \quad (26)$$

By letting

$$m \to \infty, \frac{m}{N} \to 0, \quad (27)$$

this estimator also becomes consistent (Vasicek (1976), Beirlant et al. (1997)). 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 values, gets smaller and smaller. This behavior is illustrated in Figure 1. Each plot shows, for a different value of m, the distribution of the ratio between random m-spacings and their expected value. The upper left plot shows that for $m = 1$, this distribution is distributed very widely. As m grows, the probability mass for each m-spacing concentrates around its expected value.

Such plots, which are functions of the probablity mass of intervals defined by order statistics, have the same form for *all probability distributions with continuous cumulative distribution functions*. That is, the form depends only on the value of m and not at all on the probability law. This is again a consequence of the uniformity in distribution of the probability integral transform for any (impulse-free) density.

A modification of (26) in which the $m-$spacings overlap[5]:

$$\hat{H}_{RADICAL}(Z^1,...,Z^N) \equiv \frac{1}{N-m} \sum_{i=1}^{N-m} \log\left(\frac{N+1}{m}(Z^{(i+m)} - Z^{(i)})\right), \quad (28)$$

is used in RADICAL. This is equivalent asymptotically to the estimator (14) of Vasicek (1976). Weak and strong consistency have been shown by various authors under a variety of general conditions assuming (27). For details of these results, see the review paper by Beirlant et al. (1997). Perhaps the most important property of this estimator is that it is asymptotically efficient, as shown by Levit (1978).

It is interesting to remark that while (25) and (26) have a natural correspondence to density estimates (if we ignore the region outside the range of the samples), there is no trivial correspondence between (28) and a density estimate. We are thus solving the entropy estimation problem without demanding that we solve the density estimation problem[6].

We note that Pham (2000) defined an ICA contrast function as a sum of terms very similar to (26). 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 (26) 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.

5. Allowing the m-spacings to overlap reduces the asymptotic variance of the estimator.
6. For those who are skeptical that this is possible, we suggest that it is no different than estimating the variance of a random variable without estimating its density.

3. RADICAL in Two Dimensions

Given that we have an entropy estimate in hand, we now discuss its application to optimizing Equation (13). We first discuss aspects of the estimator in the context of the two-dimensional ICA problem. Later we will extend the optimization method to multiple dimensions. Two issues which arise and which must be dealt with are local minima and what we will refer to as "false minima". The first issue is intrinsic to the optimization criterion and appears difficult to address without adopting an exhaustive search strategy. The second is a function of the estimator and will motivate a smoothing approach. We address these issues by way of some canonical examples before proceeding to a more detailed discussion of the algorithm.

3.1 Canonical empirical examples

We use three canonical examples- separation of (1) two uniform densities, (2) two double-exponential densities, and (3) two bi-modal Gaussian mixture densities - in which we examine 150 equally spaced rotations of the data between 0 and 90 degrees. Each of these examples illustrates various aspects of the estimator and the optimization procedure.

Consider Figure 2 which shows some results from separating a mixture of two uniform densities. Figure 2(a) shows the results over 100 Monte Carlo trials in which $N = 250$ and $m = 16 \approx \sqrt{250}$. As can be seen, the mean estimate (over 90 degrees of rotation) is fairly smooth with a clear global maximum at 45 degrees and a minimum at 0 degrees rotation (or equivalently 90 degrees). However, not surprisingly, any one trial has several false local minima and maxima. In Figure 2(a), for example, the individual trials which exhibited the largest positive and negative deviation from the average (for any single angle θ) over all trials are overlaid on the average result. Similar figures for the other two cases are shown in Figures 3(a) and 4(a), although local minima and maxima (over individual trials)) are not as severe in these cases.

In particular, false minima can become quite severe when the sample size is small. They 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 size there exist rotations (instances of W) for which portions of the data spuriously approximately align, producing an artificial spike in one of the marginal distributions. This is most 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 ($O(N^2)$ of them, in fact) 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 in one of the empirical marginal distributions, which in turn gives this interval an average logarithm of $-\infty$. This results in a marginal 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 (see Figure 6), it must be addressed to achieve optimal performance at least while $N \leq 2000$.

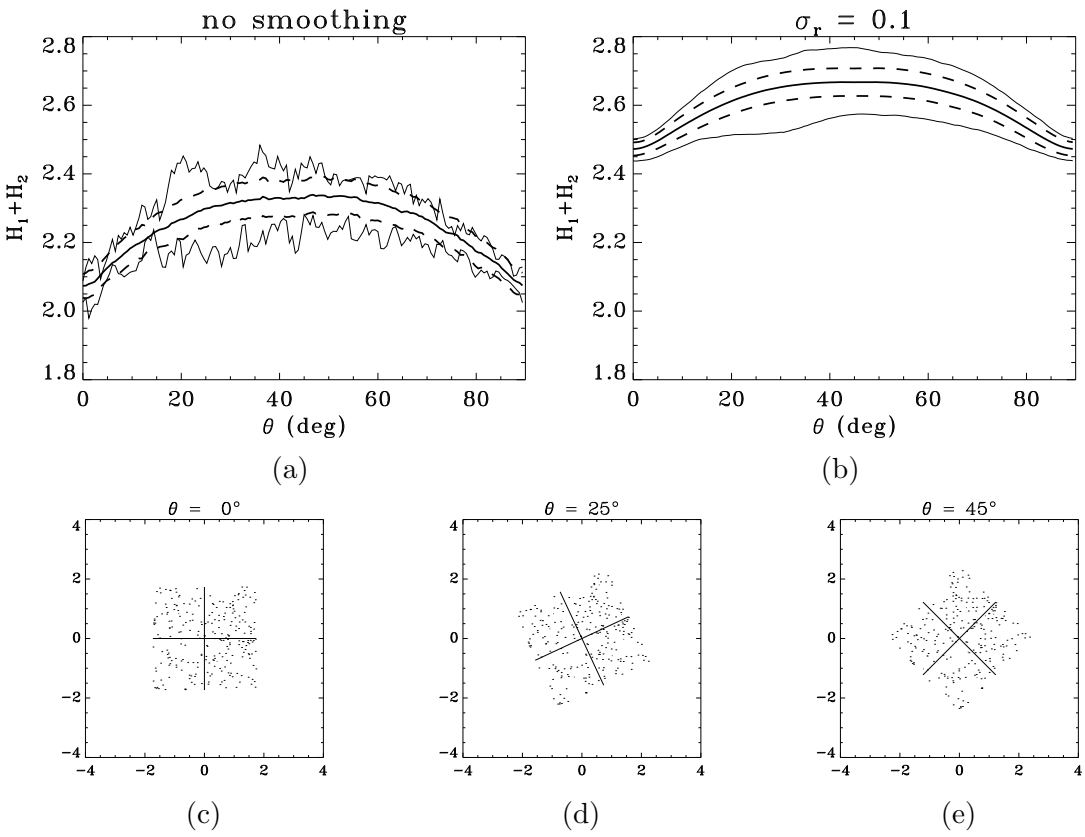


Figure 2: RADICAL for a mixture of two uniform densities: (a) The thick solid curve is the mean estimate of Equation (13) over 100 Monte Carlo trials with no data augmentation. The dotted curves indicate plus or minus one standard deviation, while the thinner (less smooth) curves are the two trials which had the largest positive and negative deviation from the mean respectively. (b) is exactly the same as (a) except that the data set was augmented with $R = 30$ and a smoothing factor of $\sigma_r = 0.1$ was used. (c) is one realization of the original sources with no rotation, (d) with 25 degrees rotation, and (e) with 45 degrees rotation. Axes of rotation are overlaid on the plots.

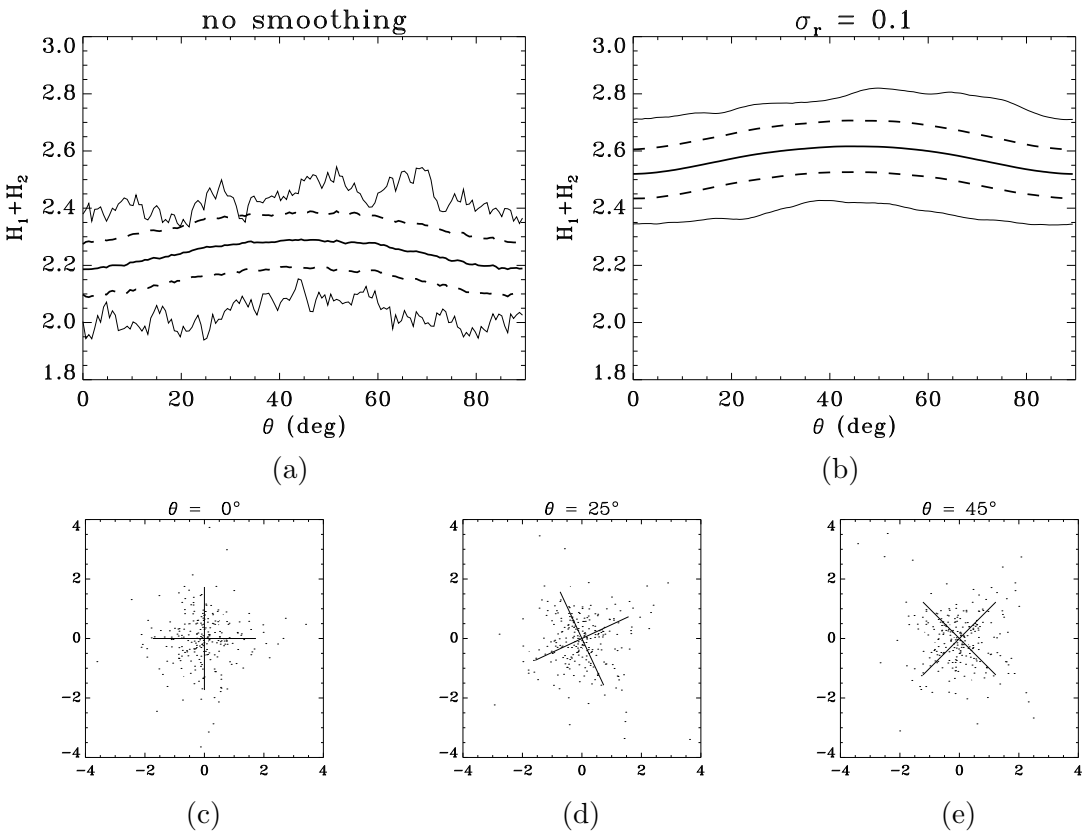


Figure 3: RADICAL for a mixture of two double-exponential densities: (a) The thick solid curve is the mean estimate of Equation (13) over 100 Monte Carlo trials with no data augmentation. The dotted curves indicate plus or minus one standard deviation, while the thinner (less smooth) curves are the two trials which had the largest positive and negative deviation from the mean respectively. (b) is exactly the same as (a) except that the data set was augmented with $R = 30$ and a smoothing factor of $\sigma_r = 0.1$ was used. (c) is one realization of the original sources with no rotation, (d) with 25 degrees rotation, and (e) with 45 degrees rotation. Axes of rotation are overlaid on the plots.

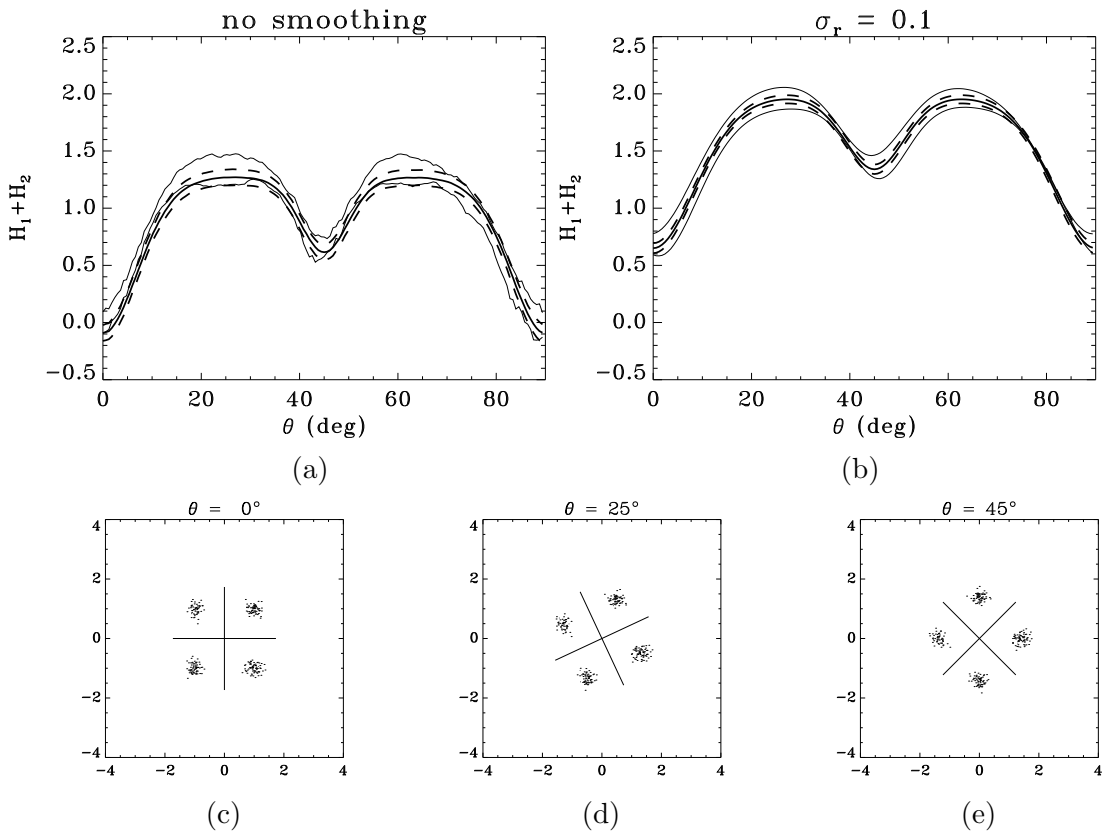


Figure 4: RADICAL for a mixture of two bi-modal densities: (a) The thick solid curve is the mean estimate of Equation (13) over 100 Monte Carlo trials with no data augmentation. The dotted curves indicate plus or minus one standard deviation, while the thinner (less smooth) curves are the two trials which had the largest positive and negative deviation from the mean respectively. (b) is exactly the same as (a) except that the data set was augmented with $R = 30$ and a smoothing factor of $\sigma_r = 0.1$ was used. (c) is one realization of the original sources with no rotation, (d) wth 25 degrees rotation, and (e) with 45 degrees rotation. Axes of rotation are overlaid on the plots.

To address this problem, 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 $N(X^j, \sigma_r^2 I)$, where R and σ_r^2 become parameters of the algorithm. We refer to this as *augmenting* the data set X. Then we use the entropy estimator (28) on the augmented data set X'. R was chosen to have a value of 30 for most of the experiments in this report. Results of this modification to the estimator are shown in Figures 2(b), 3(b) and 4(b). While not completely eliminating the problem of local minima and maxima, the results over the worst two trials are significantly smoother.

After reducing the effect of false minima on the optimization of W, we must still address legitimate local minima, the best example of which is shown in Figure 4 where a rotation of 45° is a true local minimum of the objective function. For two-dimensional source separation, taking advantage of the fact that $W(\theta)$ is a one-dimensional manifold, we do an exhaustive search over W for K values of θ. Note that we need only consider θ in the interval $[0, \frac{P_i}{2}]$, since any 90 degree rotation will result in equivalent independent components. In all of our experiments, we set $K = 150$. Importantly, it turns out that even in higher dimensions, our algorithm will remain linear in K (although polynomially more expensive in other respects), so it is relatively inexpensive to do a finer grain search over θ if desired. Complexity issues will be discussed in more detail below.

3.2 The ICA algorithm

RADICAL is a very simple algorithm. Assuming that our observed data have already been whitened, there are really only two remaining steps. The first is to generate an augmented data set X' from X by the procedure described in the previous section. The second is to minimize the cost function (13), which we do by exhaustive search. Various aspects of RADICAL are summarized in Figure 5.

There are four parameters with which we experimented informally in our early experiments. 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 guarantees the asymptotic consistency of our procedure as long as none of the marginal densities have impulses. When condition (27) is satisfied, the entropy estimator will be consistent and should perform well for large N. For small N, performance can be improved by choosing m according to the particulars of the distribution, but since the distribution is unknown in general, we avoided this and chose a fixed rule for m as a function of N for all of our experiments.

A second parameter is the number of points R used to replace each original point X^j when creating the augmented data set. The value of R can be made smaller for large N, and as N and m get large enough, point replication is entirely unnecessary, since the optimization landscape will eventually become smooth (to within the resolution of the search algorithm). However, at $N = 4000$, the largest value with which we experimented, point replication was still necessary. The experiments in this paper all used a value of $R = 30$, irrespective of the original sample size N.

Algorithm:	RADICAL, two-dimensional version.
Input:	Data vectors $X^1, X^2, ..., X^N$, assumed whitened.
Parameters:	m: Size of spacing. Usually equal to $\sqrt{N}$. σ_r^2: Noise variance for replicated points. R: Number of replicated points per original data point. 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 θ, rotate the data to this angle ($Y = W(\theta) * X'$) and evaluate cost function. 3. Output the W corresponding to the optimal θ.
Output:	W, the demixing matrix.

Figure 5: A high-level description of RADICAL for the two-dimensional ICA problem.

Next we examine σ_r^2, the variance of the R added points for each of the N points X^j. As expected, from informal testing, we found that performance was somewhat better if we allowed the variance to shrink as N grew. However, performance was relatively robust to the choice of this parameter and we chose only two different values of σ_r for all of our experiments. For $N < 1000$, we set $\sigma_r = 0.35$ and for $N >= 1000$, we halved this value, setting $\sigma_r = .175$.

The only remaining parameter for RADICAL in two dimensions is K, the number of rotations at which to measure the objective function. In informal experiments, we tried values of $50, 100, 150$, and 250. There was no noticable improvement in performance after for $K > 150$, even for $N = 4000$ and the higher dimensional tests. Of course, with N large enough, one could benefit to some extent by increasing K. Since in two dimensions, the error metric (see below) is proportional to the difference in angle between the estimated θ and the optimal θ, it is easy to see that asymptotically, the expected error is a function of K and is approximately

$$\frac{1}{2}\frac{\frac{\pi}{2}}{K} = \frac{\pi}{4K}. \tag{29}$$

For $K = 150$, this asymptotic expected error is approximately 0.005, a number small enough so that this is only a minor contribution to the total error (see experiments) for values of N considered here. 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.

3.3 Algorithmic complexity

An upper bound on the algorithmic complexity of RADICAL in two dimensions is fairly straightforward to compute. There are a number of opportunities for speedup which we will leave for future work. We will assume for this analysis and for the higher dimensional case discussed later that D, the dimension, is less than N, the sample size.

We assume that the data has been whitened to begin with. Whitening the data is $O(D^2N)$. In two dimensions, we will treat D as a constant, so this gives us $O(N)$.

Augmenting the data set with R noisy copies of each point is just $O(NR)$. Let $N' = NR$ be the size of the augmented data set. Rotation of the augmented data points to an angle θ by matrix multiplication is at most $O(D^2 N')$, but again for fixed D, we can call D^2 a constant, so this reduces to $O(N')$. In two dimensions, our estimator requires two one-dimensional sorts, which will take time $O(N' \log N')$ and two sums over at most N' spacings, which is $O(N')$. Thus, evaluating the objective function once, which involves matrix multiplication, sorting, and summing, takes time $O(N') + O(N' \log N') + O(N') = O(N' \log N')$. Note that m, the spacing size, does not enter into the complexity of evaluating the objective function.

We repeat this procedure K times in our exhaustive search. This gives us an upper bound of $O(KN' \log N')$ for the minimization of the objective function. For the whole algorithm, including whitening, we then have $O(N) + O(KN' \log N') = O(N) + O(KNR \log(NR)) = O(KNR \log(NR))$ as the final complexity for the two-dimensional algorithm. As mentioned previously, it should be possible to reduce R to 1 for large N, so technically, we can claim that RADICAL is $O(KN \log N)$. However, for moderate and low values of N, we must still choose $R > 1$, and so we include it in our complexity analysis.

3.4 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 (2002). Our tests included comparisons with FastICA (Hyvärinen and Oja (1997)), the JADE algorithm (Cardoso (1999a)), the extended Infomax algorithm (Lee et al. (1999b)), and two versions of Kernel-ICA: KCCA and KGV (Bach and Jordan (2002)).

For each of the 18 one-dimensional densities shown in Figure 6, and which were normalized to have zero mean and unit variance, 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[7]. We then measured the "difference" between the true matrix A and the W returned by the algorithm, according to the well-known criterion, due to Amari et al. (1996):

$$d(A, W) = \frac{1}{2D} \sum_{i=1}^{D} \left(\frac{\sum_{j=1}^{D} |b_{ij}|}{\max_j |b_{ij}|} - 1 \right) + \frac{1}{2D} \sum_{j=1}^{D} \left(\frac{\sum_{i=1}^{D} |b_{ij}|}{\max_i |b_{ij}|} - 1 \right), \tag{30}$$

where $b_{ij} = (AW^{-1})_{ij}$.

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

7. 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.

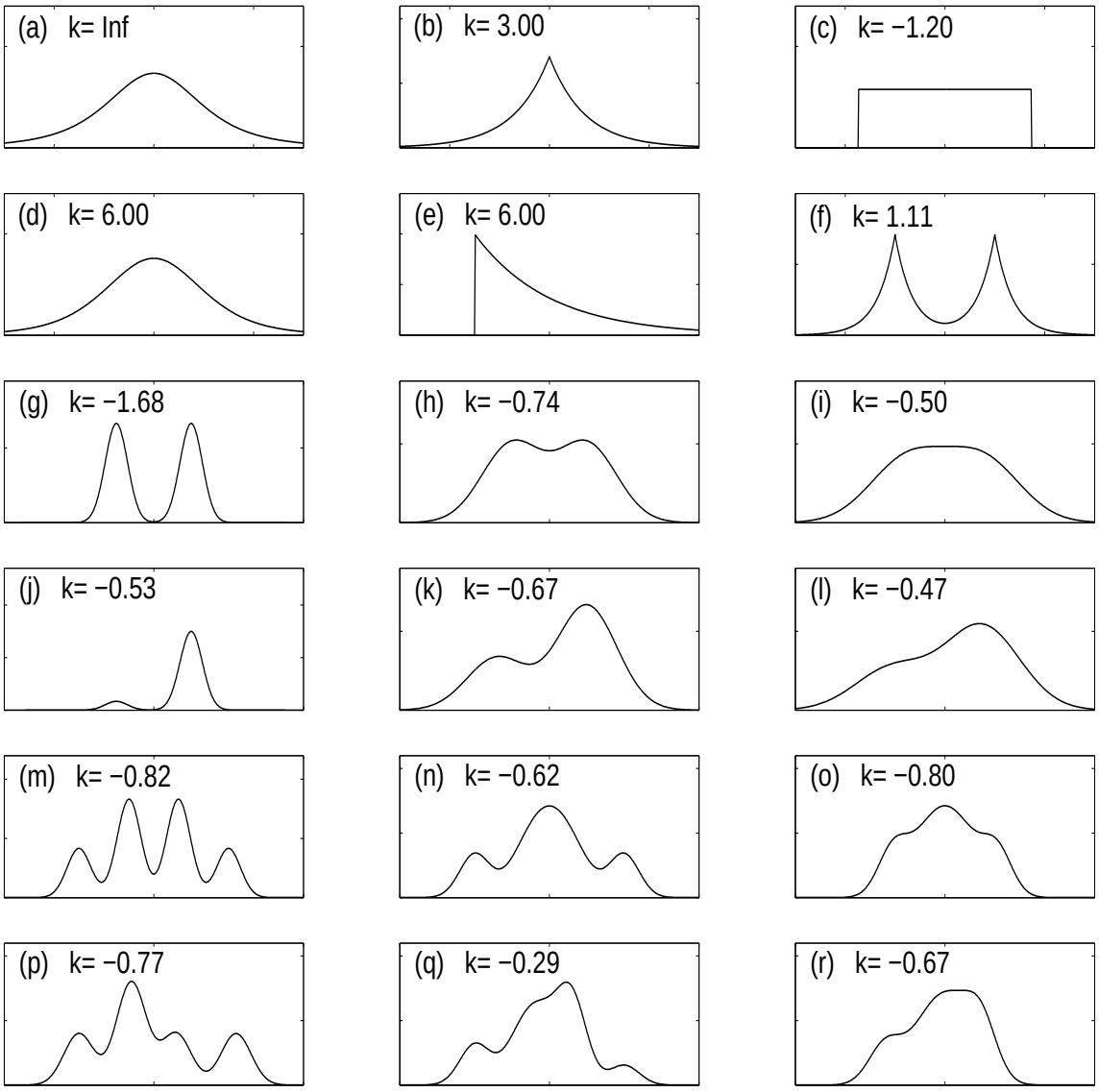


Figure 6: Probability density functions of sources with their kurtoses: (a) Student with three degrees of freedom; (b) double exponential; (c) uniform; (d) Student with five degrees of freedom; (e) exponential; (f) mixture of two double exponentials; (g)-(h)-(i) symmetric mixtures of two Gaussians: multimodal, transitional and unimodal; (m)-(n)-(o) symmetric mixtures of four Gaussians: multimodal, transitional and unimodal; (p)-(q)-(r) nonsymmetric mixtures of four Gaussians: multimodal, transitional and unimodal. This figure and code to sample from these distributions was generously provided by Francis Bach, UC Berkeley.

pdfs	FastICA	Jade	Imax	KCCA	KGV	RADICAL
a	8.9	7.5	56.3	6.6	5.7	**5.6**
b	10.2	9.3	61.8	8.4	**6.2**	7.0
c	4.4	3.1	18.4	4.7	4.3	**2.4**
d	11.8	**10.0**	61.1	13.1	11.6	12.6
e	8.1	7.4	67.7	3.7	3.1	**1.7**
f	7.9	5.5	12.4	3.6	3.3	**2.0**
g	3.9	2.9	18.1	3.1	2.9	**1.4**
h	11.1	**8.2**	27.2	10.8	8.4	12.1
i	18.5	**16.7**	37.6	25.2	23.2	27.0
j	12.2	12.8	50.5	3.1	3.0	**1.7**
k	14.1	10.3	30.2	6.1	**5.2**	5.5
l	22.6	16.4	39.2	10.6	**8.7**	11.7
m	8.2	6.9	29.5	14.8	12.3	**1.9**
n	11.4	9.7	32.1	16.3	9.7	**3.9**
o	8.7	**6.8**	23.7	13.0	9.4	8.6
p	9.9	6.7	29.1	7.7	6.0	**2.6**
q	35.8	32.0	39.1	12.4	9.4	**5.3**
r	13.0	9.5	27.7	9.7	**7.2**	8.9
mean	12.3	10.1	36.8	9.6	7.8	**6.8**
rand	10.7	8.5	29.6	8.3	6.0	**5.8**

Table 1: The Amari errors (multiplied by 100) for two-component ICA with 250 samples. For each pdf (from a to r), 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.

pdfs	FastICA	Jade	Imax	KCCA	KGV	RADICAL
a	4.4	3.7	**1.8**	3.7	3.0	2.1
b	5.8	4.1	3.4	3.7	2.9	**2.7**
c	2.3	1.9	2.0	2.7	2.4	**1.2**
d	6.4	6.1	6.9	7.1	5.7	**5.3**
e	4.9	3.9	3.2	1.7	1.5	**0.9**
f	3.6	2.7	**1.0**	1.7	1.5	**1.0**
g	1.8	1.4	**0.6**	1.5	1.4	**0.6**
h	5.1	4.1	**3.1**	4.6	3.6	3.7
i	10.0	6.8	7.8	8.3	**6.4**	8.3
j	6.0	4.5	50.6	1.4	1.3	**0.8**
k	5.8	4.4	4.2	3.2	2.8	**2.7**
l	11.0	8.3	9.4	4.9	**3.8**	4.2
m	3.9	2.8	3.9	6.2	4.7	**1.0**
n	5.3	3.9	32.1	7.1	3.0	**1.8**
o	4.4	**3.3**	4.1	6.3	4.5	3.4
p	3.7	2.9	8.2	3.6	2.8	**1.1**
q	19.0	15.3	43.3	5.2	3.6	**2.3**
r	5.8	4.3	5.9	4.1	3.7	**3.2**
mean	6.1	4.7	10.6	4.3	3.3	**2.6**
rand	5.1	4.1	6.8	3.1	**2.0**	2.1

Table 2: The Amari errors (multiplied by 100) for two-component ICA with 1000 samples. For each pdf (from a to r), averages over 100 replicates are presented. 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 from random among the 18 possible pdfs.

the set of 18 densities 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.

Table 2 shows an analogous set of results for larger data sets, with $N = 1000$. Again, RADICAL outperforms the other algorithms for most densities. However, Kernel-ICA outperforms RADICAL by a small margin in the randomized experiments.

3.5 Robustness to outliers

Figure 7 shows results for our outlier experiments. These experiments were again replications of the experiments performed by Bach and Jordan (2002). Following Bach and Jordan, we simulated outliers by randomly choosing up to 25 data points to corrupt. This was done by adding the value +5 or -5 (chosen with probability 1/2) to a single component in each of the selected data points. We performed 100 replications using source distributions chosen uniformly at random from the 18 possible sources.

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

Clearly RADICAL will be more useful if it can be applied in higher dimensions than $D = 2$. While projections and rotations of high dimensional data present no challenge, one might worry that our objective function is difficult to minimize, especially since our entropy estimator is not differentiable. It is known all too well that exhaustive search in more than a few dimensions is infeasible, as its complexity is $O(N^D)$, where N must be large to insure accuracy.

It turns out, however, that for the ICA problem, the minimization can still be approached in an "exhaustive" manner. Successive minimizations along different pairs of dimensions works well. That is, we can recast the $D-$dimensional ICA problem as a series of two-dimensional ICA problems, which we can solve well. Empirically, we show that for dimensions as high as 16, RADICAL on average outperforms or performs similarly to all other algorithms against which we tested it.

4.1 Jacobi rotations

To find the $D-$dimensional rotation matrix W^* that optimizes (13) in D dimensions, we use Jacobi methods such as those used to solve symmetric eigenvalue problems (see Golub and Loan (1996)). The basic idea is to rotate the augmented data X' two dimensions at a time using what are known as *Jacobi rotations*[8]. A Jacobi rotation of angle θ for dimensions

8. Jacobi rotations are also known as *Givens rotations*.

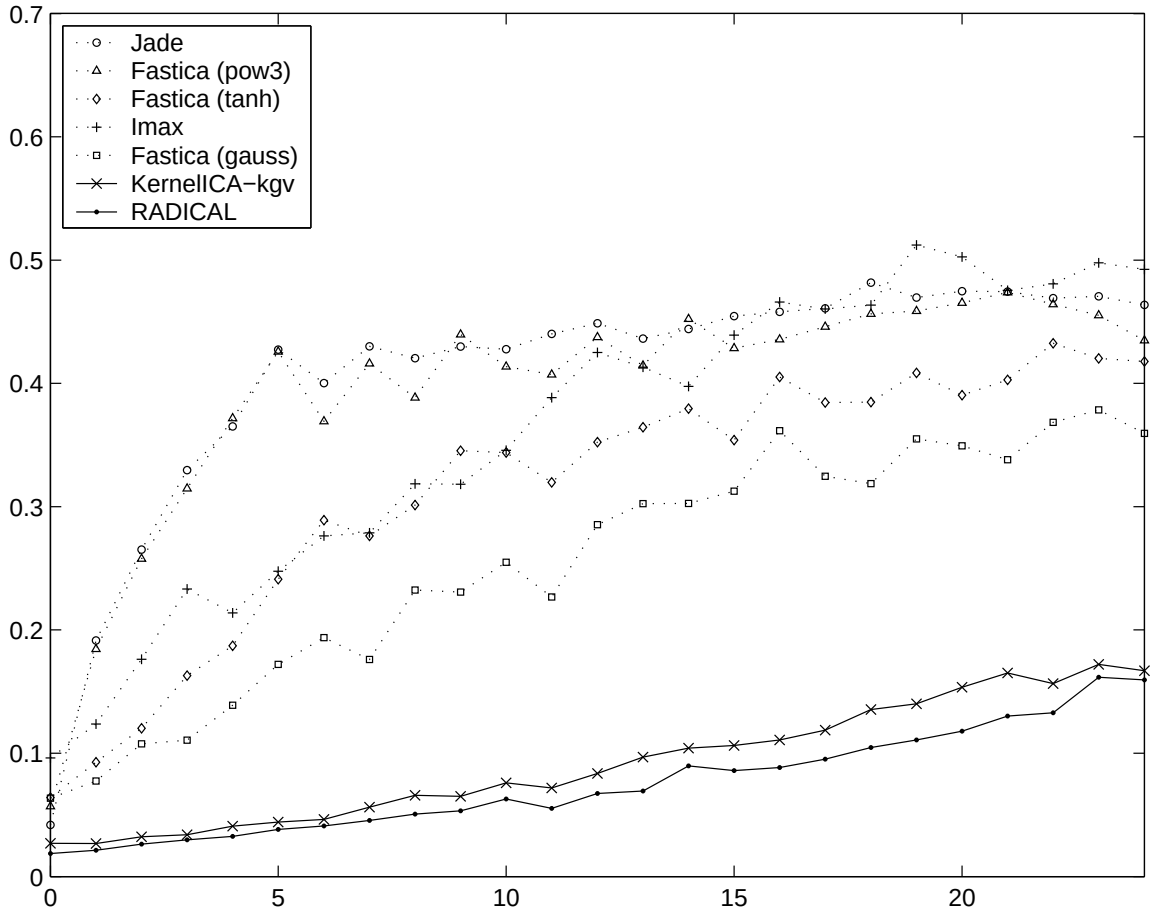


Figure 7: Robustness to outliers. The abcissa displays the number of outliers and the ordinate shows the Amari error.

p and q is defined as:

$$J(p,q,\theta) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \cos(\theta) & \cdots & -\sin(\theta) & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & \sin(\theta) & \cdots & \cos(\theta) & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}, \tag{31}$$

where the sines and cosines appear on the pth and qth rows and columns of the matrix. Since a Jacobi rotation leaves all components of an D-dimensional data point X^j unchanged except for the pth and qth components, optimizing our objective function (13) reduces to a two-dimensional ICA problem for each distinct Jacobi rotation.

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:

$$Y_{new} = J(p, q, \theta^*)Y, \tag{32}$$

keeping track of our cumulative rotation matrix:

$$W_{new} = J(p, q, \theta^*)W. \tag{33}$$

Note that since each Jacobi rotation affects only two components of Y, this is an $O(2^2NR) = O(NR)$ operation. Thus, full scale $D-$dimensional rotations need never be done (all at once). This is discussed further below.

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 $S \approx D$ provided good results for all of our multi-dimensional experiments.

4.2 Complexity of RADICAL in D dimensions

The complexity of RADICAL in D dimensions is again straightforward. Starting with the whitening of the data, we must now spend $O(D^2N)$ time on this step. Of course, we can no longer legitimately treat the dimension D as a constant. Producing the augmented data set X' now becomes an $O(DNR)$ procedure, but this will be dominated by other terms.

A single sweep through $D(D-1)/2$ Jacobi rotations produces a complexity increase by a factor of $O(D^2)$ for a total sweep complexity of $O(K(D^2)N' \log N')$. Recall that $N' = NR$ is the size of the augmented data set. The number of sweeps necessary for convergence is difficult to predict, but it seldom exceeded the dimension D. Including the number of sweeps in the complexity gives $O(SK(D^2)N' \log N')$.

It should be pointed out that this complexity analysis includes the optimization, so it should be compared against the *total run time* of other algorithms, not simply the time to evaluate the objective function.

Algorithm:	RADICAL, D-dimensional version.
Input:	Data vectors $X^1, X^2, ..., X^N$, assumed whitened.
Parameters:	m: Size of spacing. Usually equal to $\sqrt{N}$. σ_r^2: Noise variance for replicated points. R: Number of replicated points per original data point. K: Number of angles at which to evaluate cost function. S: Number of sweeps for Jacobi rotations.
Procedure:	1. Create X' by replicating R points with Gaussian noise for each original point. 2. For each of S sweeps (or until convergence): a. For each of $D(D-1)/2$ Jacobi rotations for dimensions (p, q): i. Perform 2-D RADICAL optimization, returning optimal $J(p, q, \theta^*)$. ii. Update Y according to $Y_{new} = J(p, q, \theta^*)Y$. iii. Update W according to $W_{new} = J(p, q, \theta^*)W$. 3. Output final W.
Output:	W

Figure 8: A high-level description of RADICAL for D dimensions.

dims	N	#repl	FastICA	Jade	Imax	KGV	RADICAL
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 3: Results for experiments in higher dimensions. 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 experiment replications performed to obtain the mean values at right is given in the third column. KGV is Kernel-ICA using the kernel generalized variance. Note that RADICAL performed best or equal to best in all but one experiment.

4.3 Results for D dimensions

Table 3 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. Again this data was randomly rotated, and the task was to recover the independent components. Notice that RADICAL is either the best or second best performer in each case, performing better than all other algorithms in four of seven experiments.

5. Conclusions

We have presented a novel algorithm, RADICAL, for independent component analysis. Our approach was predicated on several principles. First, direct estimation of entropy obviates the need for density estimation as an intermediate step. Second, over the space of smooth densities there are unavoidable local minima in the commonly used K-L divergence based optimization landscape. This necessitated in some respects a global search over the parameter space in order to achieve good convergence properties over a broad set of source densities. Toward that end we employed a variant of the nonparametric entropy estimator of Vasicek (1976) which is both computationally efficient and robust. In addition, our algorithm is easily used in higher dimensions. Empirical results were reported for a significant number of 2-D separation problems, 2-D separation with outliers, and a range of multi-dimensional separation problems. Our empirical results demonstrated comparable or superior results (as measured by the Amari error) to a large number of well known algorithms.

While these initial results are promising, there is still room for improvement in the algorithms as presented from both a computational and theoretical perspective. On the computational side, we take no advantage of local changes in sorting order due to local changes in rotation. Consequently, the application of standard sorting algorithms for such scenarios would be expected to greatly reduce the computational complexity of the analysis. From the theoretical perspective we presented a smoothed variant of the Vasicek estimator. Smoothing was accomplished via Monte Carlo techniques which might be avoided entirely (also reducing the computational complexity) by considering alternative methods for biasing the entropy estimate for smooth densities. Such will be the focus of our future efforts.

Acknowledgments

This work was greatly expedited and facilitated by the generous contributions of Francis Bach. He contributed a large amount of code for experiments, and provided invaluable support in replicating the excellent set of experiments developed in his previous work with Mike Jordan. We thank both of them for their help. We would also like to thank Bin Yu, David Brillinger, Andrew Ng, Jaimyoung Kwon, and Jon McAuliffe for helpful discussions.

We would like to acknowledge support for this project from ONR grant N00014-01-1-0890 under the MURI program.

Bibliography

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, 8*. Cambridge, MA: MIT Press, 1996.

Barry C. Arnold, N. Balakrishnan, and H.N. Nagaraja. *A First Course in Order Statistics*. John Wiley & Sons, 1992.

Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Math. Stat. Sci.*, 6(1):17–39, June 1997.

A. Bell and T. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

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*, 25(3):385–99, 1980.

J.-F. Cardoso. High-order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192, 1999a.

Jean-Francois Cardoso. High-order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192, 1999b.

Jean-François Cardoso and Antoine Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 17(1):161–164, January 1996. ISSN 0895-4798 (print), 1095-7162 (electronic). URL `http://epubs.siam.org/sam-bin/dbq/article/25954`.

Pierre Comon. Independent component analysis - a new concept? *Signal Processing*, 36: 287–314, 1994.

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

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

P. Hall. Limit theorems for sums of general functions of m-spacings. *Math. Proc. Camb. Phil. Soc.*, 96:517–532, 1984.

A. Hyvärinen and E. Oja. A fast fixed point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

Aapo Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Information Processing Systems 10*, pages 273–279, 1997.

Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, May 1999.

Aapo Hyvärinen. Blind source separation by nonstationarity of variance: A cumulant-based approach. *IEEE Transactions on Neural Networks*, 12(6):1471–1474, Nov 2001.

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

Solomon Kullback. *Information Theory and Statistics*. John Wiley and Sons, New York, 1959.

T. Lee, M. Girolami, A. Bell, and T. Sejnowski. A unifying information-theoretic framework for independent component analysis. *International Journal on Mathematical and Computer Modeling*, 1999a.

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*, 11(2):417–441, 1999b.

B. Ya Levit. Asymptotically optimal estimation of nonlinear functionals. *Problems of Information Transmission*, 14:65–72, 1978.

Edward B. Manoukian. *Modern Concepts and Theorems of Mathematical Statistics*. New York: Springer-Verlag, 1986.

B.A. Pearlmutter and L.C. Parra. A context-sensitive generalization of ica. In *International Conference on Neural Information Processing*, Hong Kong, Sep 1996.

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

D.-T. Pham, P. Garrat, and C. Jutten. Separation of a mixture of independent sources through a maximum likelihood approach. In *Proc. EUSIPCO*, pages pages 771–774, 1992.

Claude E. Shannon. The mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423,623–656, Jul,Oct 1948.

Oldrich Vasicek. A test for normality based on sample entropy. *Journal of the Royal Statistical Society, Series B*, 38(1):54–59, 1976.