Entropy **2008**, *10*, 745-756; DOI: 10.3390/e10040745

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

An Assessment of Hermite Function Based Approximations of Mutual Information Applied to Independent Component Analysis

Julian Sorensen

C3ID, DSTO, PO Box, Edinburgh, SA 5111, Australia, E-mail: Julian.Sorensen@dsto.defence.gov.au

Received: 23 May 2008 / Accepted: 28 November 2008 / Published: 4 December 2008

Abstract: At the heart of many ICA techniques is a nonparametric estimate of an information measure, usually via nonparametric density estimation, for example, kernel density estimation. While not as popular as kernel density estimators, orthogonal functions can be used for nonparametric density estimation (via a truncated series expansion whose coefficients are calculated from the observed data). While such estimators do not necessarily yield a valid density, which kernel density estimators do, they are faster to calculate than kernel density estimators, in particular for a modified version of Renyi's entropy of order 2. In this paper, we compare the performance of ICA using Hermite series based estimates of Shannon's and Renyi's mutual information, to that of Gaussian kernel based estimates. The comparisons also include ICA using the RADICAL estimate of Shannon's entropy and a FastICA estimate of neg-entropy.

Keywords: ICA, nonparametric estimation, Hermite functions, kernel density estimation

1. Introduction

Many of the techniques used to perform *Independent Component Analysis* (ICA) minimize an "information" measure. One of the pioneering papers for ICA [4], uses the Gram-Charlier expansion to give an approximation of differential entropy in terms of third and fourth order cumulants; FastICA [7] approximates neg-entropy, which is a measure of non-gaussianity, with a number of non-linear functions; RADICAL [10] uses order statistics to estimate differential entropy; in [3] kernel density estimators are used to estimate differential entropy; in [6] kernel density estimators are used to estimate Renyi's mutual information; in [5], nonparametric density estimation using Legendre polynomials are used to

estimate Renyi's mutual information. At the heart of these works is a nonparametric density estimator, which is then used to approximate the information measure. In this paper we will be comparing the performance of some existing ICA techniques whose basis is a nonparametric estimator of Shannon's or Renyi's mutual information, to that of performing ICA by nonparametric estimation of Shannon's and Renyi's mutual information by Hermite functions.

This paper is organized as follows. Section 2 introduces the ICA problem and the usefulness of Shannon's and Renyi's mutual information for solving it; Section 2.1 outlines how orthogonal functions can be used for nonparametric density estimation, and defines the Hermite polynomials and functions; Section 2.2 outlines kernel density estimation, which will be used in the ICA comparisons; Section 2.3 outlines the justification for the differential entropy estimate used in the RADICAL algorithm, which will be used in the ICA comparisons; Section 2.4 outlines a neg-entropy estimate used by FastICA, which will be used in the ICA comparisons. Section 3 discusses the results of a simulation experiment comparing six approaches to ICA. Section 4 discusses how the ideas in this paper could be extended to the complex-valued case.

2. Independent Component Analysis and Mutual Information

The term *Independent Component Analysis* (ICA) refers to a range of techniques that have emerged since the mid-nineties to tackle the following blind signal separation problem: find the unknown mixing matrix $\mathbf{A}$ in the following equation with only a knowledge of $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \dots & y_L \end{bmatrix}^T$

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

where the components of $\mathbf{x}$ are independent. In [4], it is shown that if the components of $\mathbf{x}$ are independent and at most one Gaussian, then the matrix minimizing Shannon's mutual information uniquely identifies $\mathbf{A}^{-1}$ up to a permutation matrix (acting on rows); Shannon's mutual information is defined for a random vector $\mathbf{Z} = \begin{bmatrix} Z_1 & Z_2 & \dots & Z_L \end{bmatrix}^T$ as

$$\begin{aligned} I(\mathbf{Z}) &:= \mathrm{E}\left[\log \frac{p_{\mathbf{Z}}(\mathbf{z})}{\prod_{i=1}^{L} p_{Z_i}(z_i)}\right] = \sum_{i=1}^{L} H(Z_i) - H(\mathbf{Z}) \\ &:= -\sum_{i=1}^{L} \int_{\mathbb{R}} p_{Z_i}(z_i) \log p_{Z_i}(z_i)\, dz_i + \int_{\mathbb{R}^L} p_{\mathbf{Z}}(\mathbf{z}) \log p_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z} \,, \end{aligned} \tag{2}$$

where $H(\cdot)$ is called differential entropy. The following quantity need not be calculated

$$\int_{\mathbb{R}^L} p_{\mathbf{Z}}(\mathbf{z}) \log p_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z} \,,$$

due to the following relation

$$\begin{aligned} I(\mathbf{BY}) &= \sum_{i=1}^{L} H(\mathbf{b}_i\mathbf{Y}) - H(\mathbf{BY}) = \sum_{i=1}^{L} H(\mathbf{b}_i\mathbf{Y}) - H(\mathbf{BAX}) \\ &= \sum_{i=1}^{L} H(\mathbf{b}_i\mathbf{Y}) - \log|\det(\mathbf{BA})| - H(\mathbf{X}) \,, \end{aligned}$$

where $\mathbf{b}_i$ is the i^{th} row of $\mathbf{B}$. This observation implies that minimizing $I(\mathbf{BY})$ is equivalent to minimizing

$$\sum_{i=1}^{L} H(\mathbf{b}_i\mathbf{Y}) - \log|\det(\mathbf{B})| \ . \tag{3}$$

This still requires that $\sum_{i=1}^{L} H(\mathbf{b}_i\mathbf{Y})$ be minimized, but this is much easier than calculating the multivariate version of differential entropy. Most papers calculate $H(\mathbf{b}_i\mathbf{Y})$ in some sort of nonparametric fashion as knowledge of the source distributions is generally unknown. For example, in [4], the Gram-Charlier expansion is used to estimate the densities, which with some approximations gave an approximation of differential entropy in terms of third and fourth order cumulants.

Most ICA methods that attempt to minimize mutual information start by whitening the data, which is useful in the two ways. Suppose $\mathbf{E}$ is a matrix that whitens $\mathbf{Y}$; let

$$\mathbf{Z} = \mathbf{EY} = \mathbf{EAX} = \mathbf{CX} \ ,$$

where $\mathbf{C} = \mathbf{EA}$, which turns out to be orthogonal, see[12]. The orthogonality of $\mathbf{C}$ is useful in two ways. Firstly, the search for the inverse mixing matrix is restricted to orthogonal matrices. Secondly, if $\mathbf{D}$ is an orthogonal matrix, then $I(\mathbf{DZ})$ simplifies to

$$I(\mathbf{DZ}) \ = \ \sum_{i=1}^{L} H(\mathbf{d}_i\mathbf{Z}) - \log|\det(\mathbf{DC})| - H(\mathbf{X}) = \sum_{i=1}^{L} H(\mathbf{d}_i\mathbf{Z}) - H(\mathbf{X}) \ , \tag{4}$$

where $\mathbf{d}_i$ is the i^{th} row of $\mathbf{D}$. This observation implies that minimizing $I(\mathbf{DZ})$ is equivalent to minimizing

$$J = \sum_{i=1}^{L} H(\mathbf{d}_i\mathbf{Z}) \ . \tag{5}$$

Shannon's mutual information is not the only information measure used for ICA. In [5] and [6], a modified version of Renyi's mutual information is used. Renyi's mutual information for a random vector $\mathbf{Z}$ is given by

$$I_\alpha(\mathbf{Z}) \ = \ \frac{1}{1-\alpha}\log\int_{\mathbb{R}^L} \frac{p_{\mathbf{Z}}(\mathbf{z})^\alpha}{\prod_{i=1}^{L} p_{Z_i}(z_i)^{\alpha-1}}\, d\mathbf{z} \ . \tag{6}$$

This quantity is minimized if the components of $\mathbf{Z}$ are independent. Unfortunately, unlike Shannon's mutual information in equation (2), Renyi's mutual information does not split up into one part involving marginal distributions and another the joint distribution. To overcome this, they instead use

$$\sum_{i=1}^{L} H_\alpha(Z_i) - H_\alpha(\mathbf{Z}) \ := \ -\frac{1}{1-\alpha}\sum_{i=1}^{L}\log\int_{\mathbb{R}} p_{Z_i}(z_i)^\alpha\, dz_i + \frac{1}{1-\alpha}\log\int_{\mathbb{R}^L} p_{\mathbf{Z}}(\mathbf{Z})^\alpha\, d\mathbf{z} \ . \tag{7}$$

This quantity is minimized if the components of $\mathbf{Z}$ are independent. Moreover, the property of Shannon's mutual information displayed in equation (4), is replicated in this measure. Hence, in [5] and [6], they use whitened data and the following quantity

$$J_\alpha = \sum_{i=1}^{L} H_\alpha(Y_i) \ . \tag{8}$$

In the following subsections we will describe some ways of estimating the marginal entropies in equation (5) and (8).

2.1. Orthogonal Series for Nonparametric Density Estimation

This material comes from [17]. Given a random sample $\{x_i\}_{i=1}^N$ for a random variable X with density $p(x)$ and a set of basis functions $\{\xi_n(x)\}_{n=0}^\infty$ which are orthogonal with respect to a kernel $K(x)$, an estimate of $p(x)$ is

$$\hat{p}(x) = \sum_{n=0}^{M} \hat{a}_n \xi_n(x) \ ,$$

where

$$\hat{a}_n = \frac{1}{\lambda_n N} \sum_{i=1}^{N} K(x_i)\xi_n(x_i) \ ,$$

where

$$\lambda_n = \int_{\mathbb{R}} K(x)\xi_n(x)^2 \, dx \ .$$

A variety of criteria exist for choosing M, see [9] for a review (we do not discuss this issue here as it turned out not to be important). Unfortunately this basis expansion density estimate, may not be a proper density in terms on non-negativity and summation to 1 - unlike kernel density estimators. However, they are much faster to calculate than kernel density estimators for large N.

Perhaps the most commonly used basis functions for nonparametric density estimation are the Hermite functions, which are defined as

$$h_n(x) = \frac{e^{-x^2/2}}{\sqrt{2^n n! \sqrt{\pi}}} H_n(x) \ , \quad n = 0, 1, 2, \dots \ ,$$

where the $H_n(x)$ are the Hermite polynomials, which are defined as

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2} \ , \quad n = 0, 1, 2, \dots \ .$$

Note that the Hermite polynomials are orthogonal with respect to the kernel e^{-x^2}, with eigenvalues $2^n n! \sqrt{\pi}$, which means that the Hermite functions are orthogonal with respect to the kernel 1, with eigenvalues 1. Now, to the best of our knowledge, we do not whether it is better to use Hermite polynomials or Hermite functions for nonparametric density estimation. However, the Hermite functions are more commonly used. One reason may be the ease with which it is possible to prove results regarding the convergence of such an estimation; see [15] and [16] for example with regard to mean square and integrated mean square rates of convergence. We used the Hermite functions for two reasons: firstly, the tails of estimated density behaved much better using the Hermite functions; secondly, a simple expression for Renyi's entropy can be derived.

Using the Hermite functions for nonparametric density estimation gave the following estimate of a density $p(x)$

$$\hat{p}(x) = \sum_{n=0}^{M} \hat{b}_n h_n(x) \ ,$$

where

$$\hat{b}_n = \frac{1}{N} \sum_{i=1}^{N} h_n(x_i) \ .$$

In our ICA experiments, we estimate the quantity in equation (5) using the following estimate for the marginal entropies

$$-\frac{1}{N}\sum_{i=1}^{N}\log\left|\sum_{n=0}^{M}\hat{b}_n h_n(x_i)\right| . \tag{9}$$

The absolute value is taken to handle negative values. We took $M = 40$ because for $M \geq 40$, there was little difference in the performance of ICA based on mutual information using this approximation to the marginal entropies.

As previously mentioned, in [5], Legendre polynomials are used to perform ICA by minimizing an estimate of the quantity in (8). We will also perform ICA by minimizing an estimate of the quantity in equation (8) for $\alpha = 2$ using Hermite functions instead. In this case, the quantity is very fast to calculate as the marginal Renyi entropy (for $\alpha = 2$ that is) has form

$$-\log\sum_{n=0}^{M}\hat{b}_n^2 . \tag{10}$$

Again, we took $M = 40$ because for $M \geq 40$, there was little difference in the performance of ICA based on mutual information using this approximation to the marginal entropies.

We finish this subsection by presenting the following property of Hermite polynomials that was used to calculate the quantities in equations (9) and (10) efficiently:

$$H_n(x+c) = \sum_{m=0}^{n}\frac{n!}{(n-m)!m!}H_{n-m}(c)(2x)^m .$$

The last line of the following equation shows how the property was used to calculate the coefficients in equations (9) and (10)

$$\begin{aligned}\hat{b}_n &= \frac{1}{N}\sum_{i=1}^{N}h_n(x_i) = \frac{1}{N}\sum_{i=1}^{N}\frac{e^{-x_i^2/2}}{\sqrt{2^n n!\sqrt{\pi}}}\left(\sum_{m=0}^{n}\frac{n!}{(n-m)!m!}H_{n-m}(0)(2x_i)^m\right)\\ &= \sqrt{\frac{n!}{2^n\sqrt{\pi}}}\sum_{m=0}^{n}\frac{H_{n-m}(0)}{(n-m)!}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{e^{-x_i^2/2}(2x_i)^m}{m!}\right) .\end{aligned}$$

The quantity in equation (9) is calculated in a similar fashion.

2.2. *Kernel Density Estimation*

A popular method for nonparametric density estimation, called kernel density estimation, was introduced by Parzen in [13], which is as follows. Given a random sample $\{x_i\}_{i=1}^N$ for a uni-variate random variable X, the kernel density estimate of $p_X(x)$ is

$$\hat{p}_X(x) = \frac{1}{Nh}\sum_{i=1}^{N}K\left(\frac{x-x_i}{h}\right) ,$$

where $K(\cdot)$ is called the kernel and h the smoothing parameter or bandwidth (note that for a multivariate random variable, almost the same is done, with the smoothing parameter being a matrix instead).

A variety of criteria exist for choosing the bandwidth h, see [14], but we do not discuss this issue here. Unsurprisingly, the Gaussian kernel is the most commonly used, which has form

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \; .$$

In [3], ICA is done by minimizing an estimate of the quantity in equation (5) - they use Gaussian kernel density estimation, which leads them to using the following marginal entropy estimate

$$-\frac{1}{N}\sum_{j=1}^{N} \log\left(\frac{1}{hN}\sum_{i=1}^{N} \exp\left(-\frac{(x_i - x_j)^2}{2h^2}\right)\right) \; . \tag{11}$$

This estimate involves a double summation, compared to a single summation in equation (9), making it much slower to calculate for large N.

In [6], ICA is performed by minimizing an estimate of the quantity in equation (8) for $\alpha = 2$. They use a Gaussian kernel to estimate the marginal densities; the marginal Renyi entropies are given exactly by

$$-\log\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \exp\left(-\frac{(x_i - x_j)^2}{4h^2}\right)\right) \; . \tag{12}$$

Again, a double summation is involved in the calculation, compared to a single summation in equation (10).

In our comparisons we will include estimates of Shannon's and Renyi's mutual information using equation (11) and (12) respectively. As is done in [3], we use an "optimal" value for the kernel bandwidth, $h = 1.06N^{-1/5}$.

2.3. RADICAL

The ICA algorithm RADICAL, proposed in [10], uses an estimate of differential entropy based on order statistics. In our comparisons, ICA will also be performed by using this estimate for the marginal entropies in equation (5). Their estimate of differential entropy is derived as follows. Given a random sample $Z_1, Z_2, \ldots, Z_N$ for a random variable Z, let $Z^{(1)} \leq Z^{(2)} \leq \ldots \leq Z^{(N)}$ denote the sample sorted in ascending order. Let $p(\cdot)$ and $P(\cdot)$ be the probability density function (what we have been referring to as density) and cumulative density functions for Z respectively. Then

$$\mathrm{E}\left[P(Z^{(i+1)}) - P(Z^{(i)})\right] = \frac{1}{N+1} \; .$$

This leads to the following nonparametric estimate for $p(\cdot)$

$$\hat{p}(z; Z_1, \ldots, Z_N) = \frac{\frac{1}{N+1}}{Z^{(i+1)} - Z^{(i)}} \; , \quad Z^{(i)} \leq z < Z^{(i+1)} \; ,$$

where $Z^{(0)}$ and $Z^{(N+1)}$ are infimum and supremum of the support of $p(z)$ respectively. A corresponding estimate of entropy based on this estimate of $p(z)$ is

$$H(Z) \;\; = \;\; -\int_{-\infty}^{\infty} p(z) \log p(z) \; dz$$

Table 1. Distributions used in Scenarios

	Distribution 1	Distribution 2
Scenario 1	$I_{[0,0.6]}(U)N(1,1) + I_{[0.6,1]}(U)N(-2,1)$,	$I_{[0,0.2]}(U)N(2,1) + I_{[0.2,1]}(U)N(-1.5,1)$
Scenario 2	Laplacian	Gaussian
Scenario 3	t-distribution, degree 3	t-distribution, degree 5
Scenario 4	Exponential distribution	t-distribution, degree 3
Scenario 5	Exponential distribution	Uniform
Scenario 6	Chi-squared, degree 3	Uniform
Scenario 7	Gaussian	Uniform
Scenario 8	Gaussian	t-distribution, degree 4

$$\begin{aligned} &\approx -\int_{-\infty}^{\infty} \hat{p}(z) \log \hat{p}(z)\, dz \\ &\approx \frac{1}{N-1} \sum_{i=1}^{N-1} \log(N+1)(Z^{(i+1)} - Z^{(i)}) \,. \end{aligned}$$

This estimate has a high variance; the variance is lowered by using larger spacing, which leads to their estimate of differential entropy

$$\hat{H}_{RADICAL}(Z_1, \ldots, Z_N) := \frac{1}{N-m} \sum_{i=1}^{N-m} \log\left(\frac{N+1}{m}\left(Z^{(i+m)} - Z^{(i)}\right)\right) \,. \tag{13}$$

Furthermore, they apply this to an augmented data set, where each observed point is augmented by the addition of R data points drawn form a $N(0, \sigma_r)$ distribution, with $R = 30$ and $\sigma_r = 0.1$ typically; the augmentation is done once to the mixed data, not to the "unmixed" data.

2.4. FastICA

One of the most popular methods for ICA is FastICA [7], for which there are a number of variants. In our comparisons we will include one of their measures of neg-entropy to estimate the marginal densities in equation (5). The neg-entropy for a random variable X is defined to be

$$J(X) = H(Z) - H(X) \,,$$

where $H(Z)$ is the differential entropy of a Gaussian random variable Z with the same mean and variance as X. Neg-entropy is useful for two reasons. The first reason is that it measures the "non-Gaussianity" of a random variable; the more non-Gaussian a random variable is, the more interesting. These "interesting" components can be found individually, rather than collectively, as is done when minimizing mutual information. Moreover, under certain conditions, the set of "interesting" components includes the independent components. The second reason neg-entropy is useful is that it is scale invariant, unlike differential entropy. Now, we will be taking a measure of neg-entropy used by FastICA out of context. Due to the whitening of the data, $H(Z)$ will remain constant, allowing neg-entropy to be used to estimate

differential entropy. We will use the following estimate for neg-entropy, which can be found in equation (5.47) of [8]

$$k_1 \left(\frac{1}{N} \sum_{i=1}^{N} x_i \exp(-x_i^2/2) \right)^2 + k_2 \left(\left(\frac{1}{N} \sum_{i=1}^{N} |x_i| \right) - \sqrt{2/\pi} \right)^2 , \tag{14}$$

where $k_1 = 36/(8\sqrt{3} - 9)$ and $k_2 = 1/(2 - 6/\pi)$.

3. Simulations

The performance of six approaches to ICA were compared for eight different scenarios. For each scenario, two signals were generated from particular distributions and then mixed. The following six ICA approaches were compared

1. Minimizing Shannon's mutual information via equation (5), using Hermite functions to estimate the marginal differential entropies, as described in equation (9).

2. Minimizing Renyi's mutual information via equation (8), using Hermite functions to estimate the marginal differential entropies, as described in equation (10).

3. Minimizing Shannon's mutual information via equation (5), using kernel density estimation to estimate the marginal differential entropies, as described in equation (11).

4. Minimizing Renyi's mutual information via equation (8), using kernel density estimation to estimate the marginal differential entropies, as described in equation (12).

5. Minimizing Shannon's mutual information via equation (5), using the RADICAL estimate for differential entropy, as described in equation (13).

6. Minimizing Shannon's mutual information via equation (5), using the FastICA estimate for neg-entropy in equation (14).

For each scenario, 200 simulations with 1000 data points were performed. For each simulation, the mixing matrix was also randomly generated so that the results would not be dependent on the mixing matrix. The data was also whitened, and a "full" search (rather than a gradient approach) for the inverse mixing matrix done over the space of (orthogonal) rotation matrices. Each scenario used the same two distributions; Table [1] lists the distributions used, with Figure [1] displaying the distributions. Note that in the first scenario, the distributions are Gaussian Mixtures, where U is a uniformly distributed random variable on $[0, 1]$.

The performance of each algorithm was measured with the commonly used Amari's Error [1]. Amari's error is particularly useful for ICA comparisons as it is permutation invariant. We used it in the following way. Suppose $\mathbf{W}$ is an estimate of the inverse mixing matrix $\mathbf{A}^{-1}$. Amari's error was then calculated to be

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

where $b_{ij} = (\mathbf{A}^{-1}\mathbf{W}^{-1})_{ij}$.

Figure 1. Plots of Distributions Used

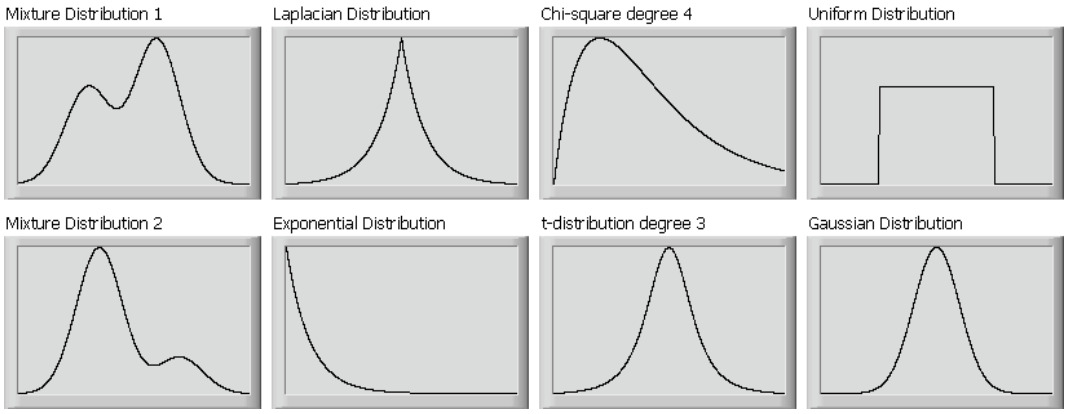

Table 2. Mean Amari's Errors (multiplied by 100)

	Shannon MI: Hermite	Renyi MI: Hermite	Shannon MI: Kernel	Renyi MI: Kernel	RADICAL	FastICA
Scenario 1	3.0	2.9	2.3	2.8	2.7	2.4
Scenario 2	5.9	4.0	3.9	3.7	5.1	3.1
Scenario 3	2.5	1.5	1.6	1.6	1.5	2.1
Scenario 4	1.9	1.3	1.4	1.4	1.3	2.2
Scenario 5	1.4	1.6	1.4	1.7	1.4	2.8
Scenario 6	1.6	2.0	1.6	2.2	1.6	2.7
Scenario 7	2.4	3.3	2.3	3.6	2.3	4.2
Scenario 8	4.9	1.8	2.1	2.2	1.7	2.2
mean	3.0	2.3	2.1	2.4	2.2	2.7

The mean Amari errors for each scenario can found in Table [2]; the last row containing the mean over all simulations. We can see that no-one method is superior or inferior to the rest. However, if we compare the mean of Amari's error over all simulations, an estimate of Shannon's mutual information based on kernel density estimators performs best. Also, using Hermite functions, ICA using Renyi's mutual information works better than Shannon's mutual information. A possible explanation for this is displayed in Figure [2]: the first graph displays the estimates of Shannon's mutual information over a range of rotation matrices; the second graph displays estimates of Renyi's mutual information for the same rotation matrices. Shannon's and Renyi's estimate of mutual information using Hermite functions displayed different behaviors, that is, Shannon's was generally not that smooth compared to Renyi's, leading to "false" minima. One possible reason for this can be seen in equation (9), where the absolute value of the estimated density is taken - the lack of smoothness in the absolute value function maybe transferred to the marginal entropy estimates.

We finish this section by comparing the time needed to calculate the various entropy estimates. One would expect that the Hermite function based estimates should be of order $O(n)$ as they really only

Figure 2. Behavior of Shannon and Renyi Mutual Information Estimates using Hermite Functions

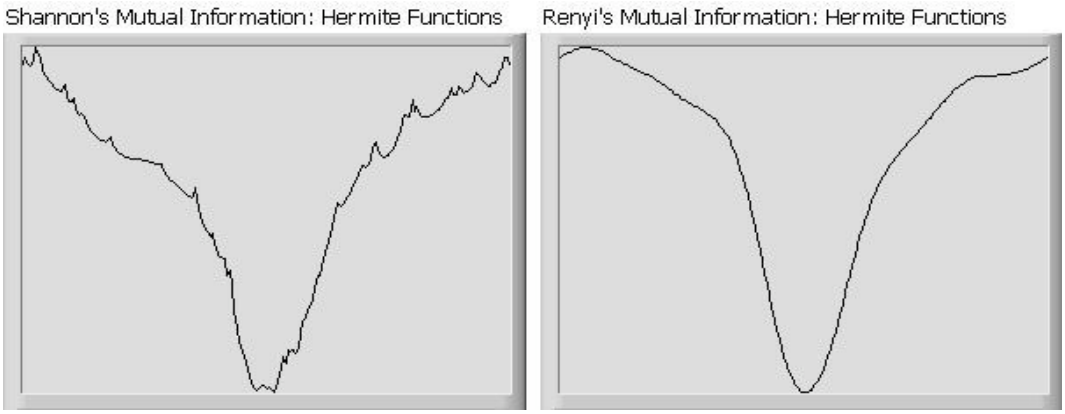

Table 3. Calculation times for entropy calculation (milliseconds)

	Shannon MI: Hermite	Renyi MI: Hermite	Shannon MI: Kernel	Renyi MI: Kernel	RADICAL	FastICA
1000 points	3	1	244	244	54	0
2000 points	5	2	485	485	110	0
4000 points	9	5	3786	3786	238	1
8000 points	21	10	14978	14978	485	2

involve one major summation; the kernel density based estimates should be of order $O(n^2)$ due to the double summation involved; the RADICAL estimate should be of order $O(n)$ as it's major component is sorting. Observations bear out these hypothesis - see Table [3], which contains the time in milliseconds taken to calculate the various entropy estimates for 1000, 2000, 4000 and 8000 data points. Moreover, we see that the Hermite based approximations are significantly faster than all but the FastICA estimate, which is to be expected as the FastICA does not involve many calculations.

4. Further Work

The complex-valued ICA case has not received the same attention as the real-valued case. Some ICA techniques have been developed for the case of sources with spherically symmetric distributions, that is, the distribution of a source is unchanged by the multiplication of $e^{i\theta}$. In this case, ICA techniques can be applied to the absolute value of un-mixed sources; both RADICAL and FastICA (see [2] and [11] respectively) have been extended to the complex case using this idea. Now, if the sources do not have spherically symmetric distributions, one cannot just consider the absolute value of the un-mixed sources. However, there is no reason why one cannot use the complex-valued versions of the information measures in equations (5) and (8). For both information measures, estimates of the entropy quantities making them up will be based on bi-variate density estimates; both Hermite series and kernel

density estimates can give these. To illustrate, consider a bi-variate random variable (X, Y) with samples $\{(x_i, y_i)\}_{i=1}^{M}$. The Hermite series estimate of the bi-variate density $p_{X,Y}(x, y)$ is

$$\hat{p}_{X,Y}(x, y) = \sum_{m=0}^{M} \sum_{n=0}^{M} \hat{b}_{n,m} h_m(x) h_n(y) \ ,$$

where

$$\hat{b}_{n,m} = \sum_{i=1}^{N} h_m(x_i) h_n(y_i) \ .$$

The bi-variate versions of the entropy estimates in equations (9) and (10) are then

$$-\frac{1}{N} \sum_{i=1}^{N} \log \left| \sum_{m=0}^{M} \sum_{n=0}^{M} \hat{b}_{n,m} h_m(x_i) h_n(y_i) \right| \ , \tag{15}$$

and

$$-\log \sum_{m=0}^{M} \sum_{n=0}^{M} \hat{b}_{m,n}^2 \ , \tag{16}$$

respectively. Nonparametric density estimation using Hermite functions does have some advantages over kernel density estimation in higher dimensions. Firstly, as was noted in [15] and [16], the Hermite series estimate converges at the same rate (with respect to mean integrated square error and mean square error) regardless of dimension, whereas the rate of convergence of the kernel density based estimate degrades as the dimension increase. The second reason is that unlike the uni-variate case, it is a non-trivial task to choose the smoothing matrix required by the kernel density estimate.

Some further work would be to compare the performance of Hermite series and kernel density based estimates of the information measures in equations (5) and (8) for the case of complex data which is not spherically symmetric. In this case, the Hermite series approach may well prove to be superior due to it's better scaling in multi-dimensions to that of kernel density estimates.

5. Conclusion

In this paper we have compared the performance of ICA using Hermite function based nonparametric estimates of Shannon's and Renyi's mutual information to that of kernel density based estimates, along with RADICAL and FastICA estimates of Shannon's mutual information. In the eight scenarios considered, the Hermite function based approach was competitive in terms of performance with the other ICA approaches, and much faster to calculate than the kernel density based approach.

Acknowledgements

The authour would like to thank Prof. B. Moran for his valuable suggestions regarding the work in this paper.

References and Notes

1. Amari, S.; Cichocki, A.; Yang, H. H. A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems*, **1996**, *8*, 757-763.

2. Bingham, E. ; Hyvarinen; A. A fast fixed-point algorithm for independent component analysis of complex-valued signals. *Int. J. of Neural Syst.* **2000**, *10 (1)* , 1-8.
3. Boscolo, R.; Pan, H.; Roychowdhury. V. P. Independent component analysis based on nonparametric density estimation. *IEEE Trans. Neural Networks.* **2004**, 55-65.
4. Como, P. Independent component analysis, A new concept? *Signal Process.* **1994**, 287-314.
5. Erdogmus, D.; Hild II, K. E.; Principe, J. C. Independent components analysis using Renyi's mutual information and Legendre density estimation. In *Proceedings of International Joint Conference on Neural Networks*, USA, July 2001, 4; pp. 2762–2767.
6. Hild II, K. E.; Erdogmus, D.; Principe, J. C. Blind source separation using Renyi's mutual information. *IEEE Signal Process. Lett.* **2001**, *8*, 174–176.
7. Hyvarinen, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Networks*. **1999**, *1*0, 626–634.
8. Hyvarinen, A.; Karhunen, J. ; Oja, E. *I*ndependent Component Analysis. John Wiley & Sons, 2001.
9. Izenman, A.J. Recent developments in nonparametric density estimation. *J. Amer. Statistical Assoc.* **1991**, *8*6, 205–223.
10. Learned-Miller, E. G. ; Fisher III, J. W. ICA using spacing estimates of Entropy. *J. Mach. Learn. Res.* **2003**, *4*, 1271–1295.
11. Lee,I. ; Lee, T. Nonparameteric Independent Component Analysis for Circular Complex Variables. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, 2007, 2; pp. 665–668.
12. Obradovic, D. ; Deco, B. Information maximization and independent component analysis: Is there a difference? *Neural Computat.* **2000**, 2085–2101.
13. Parzen, E. On esimation of a probability density function and mode. *Ann. Math. Statist.* **1962**, *3*3, 1065–1076.
14. Silverman, B. W. *D*ensity Estimation for Statistics and Data Analysis, Chapman and Hall, 1985.
15. Schwartz, S. C. Estimation of a probability density by an orthogonal series. *Ann. Math. Statist.* **1966**, *3*8, 1262–1265.
16. Walter, G. G. Properties of Hermite series estimation of probability density. *Ann. Statist.* **1977**, *5*, 1258–1264.
17. Webb, A. *Statistical Pattern Recognition*; Hodder Arnold, 1999.