The Philippine Statistician, 2007
Vol. 56, Nos. 3-4, pp. 37-51

Small Sample Statistical Properties of a Correlation Contrast Function Used in Kernel Independent Component Analysis

Roberto N. Padua and Warren A. Luzano[1]

Received: March, 2008; Revised: June, 2008

ABSTRACT

This study considers the finite-sample statistical properties of the empirical regularized maximum correlation estimator used in Kernel Independent Component Analysis through bootstrap. This study also determined the performance of Kernel ICA as evaluated by the **normalized** Amari error index. The correlation is calculated in a reproducing kernel Hilbert space. Results indicate that the variance of the computed correlation contrast function behaves as $O(n^{\lambda})$, i.e. the variance decreases at the rate n^{λ} which means that the finite sample calculations seem to indicate that the F-correlation contrast function converges in mean square to a fixed value for large n. The **normalized** Amari errors are generally of the order 10^{-3}, thus the contrast function is able to separate signals with 99.9% accuracy.

Keywords and phrases: Independent component analysis, kernel, canonical correlation, bootstrap

I. INTRODUCTION

1.1 Independent Component Analysis (ICA)

Consider a random vector $\mathbf{x} = (x_1, x_2, \ldots, x_m)$, the components of which are independent random variables. Suppose that there is an $m \times m$ mixing matrix A which mixes the random variables as follows:

$$\mathbf{y} = \mathbf{A}\mathbf{x}; \qquad (1)$$

The vector **y** is observed and the problem consists of recovering the original vector **x** from **y**. This is called the blind source separation (BSS) problem. ICA is used as a tool in solving the BSS problem. If the mixing matrix **A** were known, then the problem is a simple one consisting of inverting **A** to recover **x**. However, since **A** is unknown, given N independently, identically distributed observations of **y**, we aim to estimate **A** and recover the latent vector **x** corresponding to any particular **y** by solving a linear system. Let **W** be the inverse of A, then:

$$\mathbf{x} = \mathbf{W}\mathbf{y} \qquad (2)$$

However, the only known quantity in this expression is the vector **y** (the observed mixed signals) while all the others are unknown. The method of independent component

[1] First author is a consultant of Liceo de Cagayan University (LdCU), Cagayan de Oro City, email address: rnpadua@yahoo.com. Second author is a faculty of Misamis Oriental State College of Agriculture and Technology (MOSCAT), Misamis Oriental, and a graduating Ph.D student in Applied Mathematical Sciences of Mindanao Polytechnic State College (MPSC), Lapasan, Cagayan de Oro City, email address: stat_warren@yahoo.com.

analysis (ICA) exploits the known independence of the signals in the vector **x** in order to derive an estimate of the matrix **W** and hence, of the unknown but independent signals in **x**.

The problem is similar to the classical orthogonal Factor Analysis model in which a known *p x 1* vector **x** is decomposed into:

$$\mathbf{x} = \mathbf{L}\mathbb{F} + \boldsymbol{\varepsilon}$$

where **L** is *p x m* matrix of constants called factor loadings and $\mathbb{F}$ is an *mx1* vector of unknown factors and $\boldsymbol{\varepsilon}$ is a *px1* uncorrelated random errors. The observable random vector **x** with p components has a covariance matrix $\sum$ (where, without loss of generality, we assume that its mean vector is zero),and that **X** is linearly dependent upon m unobservable random variables $\mathbf{F}_1,\mathbf{F}_2,\ldots,\mathbb{F}_m$ called *common factors* and p additional sources of variations $\varepsilon_1, \varepsilon_2\ldots$, ε_p or *specific factors*. With so many unknowns, the classical factor analysis model assumes that:

$$E(F) = 0,\ cov(F) = I,\ E(\varepsilon) = 0 \text{ and } cov(\varepsilon) = \psi \text{ a diagonal matrix, and } cov(\varepsilon,F) = 0,$$

in order to arrive at a reasonable solution.

The solution to this factor analysis problem hinges on a reconstruction of the covariance matrix of observed vector **x** knowing that the *underlying factors are uncorrelated, the random errors are uncorrelated (with covariance matrix ψ, diagonal), and the common factor F and specific factor ε are also uncorrelated.*

Independent component analysis or ICA runs in a similar vein but instead of using the covariance structure of the known vector y, we use the *stronger requirement of independence among the original source signals.*

1.2 Independence

We are going to deal with statistical independence, as a way of characterizing signals. Independence is a statistical property. A set of signal s is said to be independent if every signal (component) gives no information on the other signals of the set. The joint probability density function is then equal to the product of marginal probability density functions. Independence implies uncorrelated signals but the converse is not true.

Technically, independence can be defined by the probability densities. Denote by $f(\mathbf{y}_1,\mathbf{y}_2)$ the joint probability density function of $\mathbf{y}_1$ and $\mathbf{y}_2$, $f(\mathbf{y}_1)$ the marginal pdf of $\mathbf{y}_1$, and similarly $f(\mathbf{y}_2)$ the marginal pdf of $\mathbf{y}_2$. Then $\mathbf{y}_1$ and $\mathbf{y}_2$ are independent if and only if the joint pdf is factorizable in the following way:

$$f(\mathbf{y}_1,\mathbf{y}_2) = f(\mathbf{y}_1)f(\mathbf{y}_2)$$

In many instances, the density function $f(\cdot)$ is not known. For this reason, it is necessary to define some other criterion for statistical independence which will not require knowledge of the underlying probability distribution $f(\cdot)$. Such criteria are called ***contrast functions.***

1.3 Contrast Functions Used in ICA

The search for acceptable contrast functions begins by considering the concept of entropy (the amount of disorder or independence). To this end, let $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ and $\mathbf{f}(\cdot)$ be a multivariate density function. The differential entropy of the vector y is given by:

$$H(\mathbf{y}) = -\int f(\mathbf{y}) \log f(\mathbf{y})\, d\mathbf{y} \tag{2}$$

Note that if y were a univariate random variable, the differential entropy reduces to the classical Shannon's diversity index (Brown, 2000). Differential entropy measures the amount of disorder present. The Gaussian variable has the largest entropy among all random variables of equal variance. In order to make (2) scale invariant, we can normalize it to give rise to the definition of negentropy.

$$J(\mathbf{y}) = H(\mathbf{y}_{gauss}) - H(\mathbf{y}) \tag{3}$$

where $H(\mathbf{y}_{gauss})$ is the entropy of the Gaussian random variable. Negentropy can also be interpreted as a measure of non-gaussianity, i.e. it is zero if **y** were a multivariate normal random variable and positive otherwise.

So far, however, all the quantities considered still involve the unknown density function $f(\cdot)$. We now attempt to derive an approximation of the negentropy function due to Jones and Sibson (1987). Negentropy is approximated using the cumulants of the distribution. To simplify the discussion, assume for the moment that y is a univariate random variable and it is centered so that is mean is zero and its variance is 1. The third cumulant, or skewness, of y can be defined as skew $(y) = E\{y^3\}$. It is a measure of the lack of symmetry of the distribution of y. The fourth cumulant, or kurtosis, of y can be defined as kurt $(y) = E\{y^4\}-3$. Using these two cumulants, the following approximation for negentropy was derived by Jones and Sibson (1987)

$$J(y) = 1/12\ \text{skew}(y)^2 + 1/48\ \text{kur}(y)^2 \tag{5}$$

For multivariate random vector **y**, the individual negentropy approximations for the individual components are computed. Now, *negentropy can be estimated without reference to the underlying probability distribution using the sample moments.*

Further, by using negentropy one can define mutual information. Mutual information is a natural measure of the dependence between the random variables. Constraining the variables to be uncorrelated, the mutual information between m (scalar) random variables can be defined as

$$I(y_1 \ldots y_m) = J(\mathbf{y}) - \sum_{i=1}^{m} J(y_i) \tag{6}$$

Since mutual information is the information-theoretic measure of non-Gaussianity, it is natural to use it as the criterion for finding the ICA transform.

A general approximation of (6) was formulated by Hyvarinen (1997). He maximizes the following function that approximates the negentropy:

$$J_G(\mathbf{W}) = [\, E\{G(y)\} - E\{G(v)\}\,]^2 \qquad (7)$$

where v is the standard Gaussian random variable. Maximizing $J_G(\mathbf{W})$ allows one to find one independent component.

1.3 Contrast Function Based on Maximum Correlation

The cumulant-based method provides a poor approximation of negentropy. They measure mainly the symmetry and the peakedness of the distribution and disregard other features of the signals.

Bach and Jordan (2003) provided a new approach to the ICA problem based not on a single nonlinear function, but on an entire function space of candidate nonlinearities. The use of a function space makes it possible to adapt to a variety of sources and thus makes the algorithm more robust to varying source distributions. They used the contrast function maximum correlation and called their ICA program as Kernel ICA.

II. THEORETICAL FRAMEWORK

2.1 Kernel Based Contrast Function

2.1.1 Feature Space and the F-Correlation

Following Bach and Jordan (2003), consider the vector space of functions of candidate nonlinearities and call this vector space F. We will call this vector space of functions the feature space. Note that skewness and kurtosis are but two features of a signal. To best estimate a signal, we need all the (infinite uncountable and countable) features of a signal. Let F be the collection of all these functions, then F contains uncountably many functions. Let us consider for simplicity two univariate random variables x_1 and x_2. Define on F the F–correlation given by:

$$\rho F = \max_{f1, f2 \in F} corr\left(f_1(x_1), f_2(x_2)\right) = \max_{f1, f2 \in F} \frac{cov\left(f_1(x_1), f_2(x_2)\right)}{\left(\mathrm{var}(f_1(x_1))^{1/2}\, \mathrm{var}(f_2(x_2))^{1/2}\right)}, \qquad (8)$$

where f_1 and f_2 ranges over F.

To find these correlations, we will utilize Canonical Correlation Analysis (CCA). CCA focuses on the correlation between the two linear combinations of x_1 and x_2. The first canonical correlation is also called the *maximum correlation.* Now, if $\mathbf{x_1}$ and $\mathbf{x_2}$ were multivariate N dimensional random vectors, then we will show later that performing CCA on these two vectors of dimension N is the same as solving *a generalized eigenvalue problem.*

The vector space (or the feature space) F is very large indeed. It allows us to identify all possible features { f_i }, distinguishable and non-distinguishable features; and correlates them, so that we are assured of independence when $\rho F = 0$.

2.1.2 The F-Correlation in Detail

Let F = { $\Phi(.)$ /Φ maps a random variable x to $\Phi(x)$, a feature of the signal}. Define the Mercer kernel K :F x F $\rightarrow$ R by the equation K(x,y) = < $\Phi(x)$/ $\Phi(y)$ >, where <./.> denotes the inner product on F x F. By the Riesz Representation Theorem on this Hilbert space, every function f ε F can be represented by f(x) = <K(.,x)/f >. Let K_1 and K_2 be Mercer kernels with feature maps Φ_1 and Φ_2 and feature spaces $\boldsymbol{F_1}$, $\boldsymbol{F_2}$. The canonical correlation between $\Phi(x_1)$ and $\Phi(x_2)$ for two random variables x_1 and x_2 is defined as:

$$\rho F = \max_{(f_1, f_2) \in F_1 x F_2} corr\left(\langle \Phi_1(x_1), f_1 \rangle, \langle \Phi_2(x_2), f_2 \rangle\right)$$

and is equal to $\rho F = \max_{(f_1, f_2) \in F_1 x F_2} corr\left(f_1(x_1), f_2(x_2)\right)$ by the above formulation of the Mercer kernels.

While in general we can use different kernels for x_1 and x_2, for notational simplicity we can also consider cases in which the two kernels and the two feature spaces are equal, denoting them as K and F respectively. Bach and Jordan(2003) *proved that if the feature space F is the reproducing kernel Hilbert space (RKHS), the canonical correlation ρ_F is zero if and only if the random variables x_1 and x_2 are independent.*

2.1.3 Kernelization of the Canonical Correlation

We have defined the F-correlation in terms of population expectation only, we need to have an empirical estimate of the F-correlation. We will call it the "kernelized" version of canonical correlation.

Let $\{x_1^1, ..., x_1^N\}$ and $\{x_2^1, ..., x_2^N\}$ denote sets of N empirical values of x_1 and x_2 respectively, and $\{\Phi(x_1^1), ..., \Phi(x_1^N)\}$ and $\{\Phi(x_2^1), ..., \Phi(x_2^N)\}$ denote the corresponding transformations in the feature space. Also, let the data be centered (that is $\sum_{k=1}^{N} \Phi(x_1^k) = \sum_{k=1}^{N} \Phi(x_2^k) = 0$). We denote the empirical canonical correlation as $\hat{\rho}F = (x_1, x_2)$, which is now based on empirical covariances and not on population covariances. But since $\hat{\rho}F = (x_1, x_2)$ depends also on the Gram matrices K_1 and K_2 of these observations, thus we can use the notation $\hat{\rho}F = (K_1, K_2)$ to denote the empirical canonical correlation.

For fixed f_1 and f_2, the empirical covariance of the projections in feature space can be written:

$$\begin{aligned} \text{cov}\left(\langle \Phi_1(x_1), f_1 \rangle, \langle \Phi_2(x_2), f_2 \rangle\right) &= \frac{1}{N} \sum_{k=1}^{N} \langle \Phi_1(x_1^k), f_1 \rangle \langle \Phi_2(x_2^k), f_2 \rangle \\ &= \frac{1}{N} \sum_{k=1}^{N} \left\langle \Phi_1(x_1^k), \sum_{i=1}^{N} \alpha_1^i \Phi(x_1^i) \right\rangle \left\langle \Phi_2(x_2^k), \sum_{j=1}^{N} \alpha_2^j \Phi(x_2^j) \right\rangle \\ &= \frac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_1^i K_1(x_1^i, x_1^k) K_2(x_2^j, x_2^k) \alpha_2^j \end{aligned}$$

$$= \frac{1}{N}\alpha_1^T K_1 K_2 \alpha_2 \tag{9}$$

where K_1 and K_2 are the Gram matrices associated with the data sets respectively.

We also obtain

$$\text{var}\left(\langle \Phi_1(x_1), f_1 \rangle\right) = \frac{1}{N}\alpha_1^T K_1 K_1 \alpha_1 \quad \text{and}$$

$$\text{var}\left(\langle \Phi_2(x_2), f_2 \rangle\right) = \frac{1}{N}\alpha_2^T K_2 K_2 \alpha_2 \tag{10}$$

Putting these results together, the kernelized CCA problem becomes that of performing the following maximization:

$$\hat{\rho}F(K_1, K_2) = \max_{\alpha_1, \alpha_2 \in R^N} \frac{\alpha_1^T K_1 K_2 \alpha_2}{\left(\left(\alpha_1^T K_1 K_1 \alpha_1\right)^{1/2}\left(\alpha_2^T K_2 K_2 \alpha_2\right)^{1/2}\right)} \tag{11}$$

Taking the derivatives with respect to the vectors α_1 and α_2 and normalizing these vectors, performing CCA on two vectors of dimension N, with covariance matrix equal to $\begin{pmatrix} K_1K_1 & K_1K_2 \\ K_2K_1 & K_2K_2 \end{pmatrix}$ is the same as solving the generalized eigenvalue problem

$$\begin{pmatrix} 0 & K_1K_2 \\ K_2K_1 & 0 \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \rho\begin{pmatrix} K_1^2 & 0 \\ 0 & K_2^2 \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

which is based on the Gram matrices K_1 and K_2. Here, we note that finding the maximum correlation is the same as finding minimal generalized eigenvalue.

2.1.4 Regularization

We need to incorporate a regularization constant to make the empirical canonical estimate a consistent estimator. The regularized F-correlation denoted by ρ_F^k is

$$\hat{\rho}F = \max_{f1, f2 \in F} \frac{cov\left(f_1(x_1), f_2(x_2)\right)}{\left(\left(\text{var}\, f_1(x_1) + \| f_1 \|_F^2\right)^{1/2}\left(\text{var}\, f_2(x_2) + \| f_2 \|_F^2\right)^{1/2}\right)}$$

where k is a small positive constant. Note that the regularized F-correlation inherits the independence characterization property of the F-correlation. In order to estimate it from a finite sample, we expand the elements inside the square root to obtain the regularized kernel CCA:

$$\hat{\rho}_F^k(K_1,K_2)=\max_{\alpha_1,\alpha_2\in R^N}\frac{\alpha_1^T K_1 K_2 \alpha_2}{\left(\left(\alpha_1^T\left(K_1+\frac{N_k}{2}I\right)^2\alpha_1\right)^{1/2}\left(\alpha_2^T\left(K_2+\frac{N_k}{2}I\right)^2\alpha_2\right)^{1/2}\right)}$$

with its equivalent formulation as generalized eigenvalue problem as follows

$$\begin{pmatrix}0 & K_1K_2\\ K_2K_1 & 0\end{pmatrix}\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}=\rho\begin{pmatrix}\left(K_1+\frac{N_k}{2}I\right)^2 & 0\\ 0 & \left(K_2+\frac{N_k}{2}I\right)^2\end{pmatrix}\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}$$

2.1.5 Free Parameters

The kernel ICA has two free parameters: the regularization parameter k and the width σ of the kernel (assuming identical Gaussian kernels for each source). It was found out that setting the parameters $k = 2 \times 10^{-3}$ and $\sigma = ½$ for large samples ($N > 1000$) and $k = 2 \times 10^{-2}$ and $\sigma = 1$ for smaller samples ($N < 1000$), the kernel ICA is relatively robust.

2.2 The Bootstrap

In many signal and information processing applications one is interested in forming estimates of a certain number of unknown parameters of a random process, using a set of sample values. Further, one is interested in finding the sampling distribution of the estimators, so that the respective means, variances, and cumulants can be calculated, or in making some kind of probability statements with respect to unknown true values of the parameters. For example one could be interested in assigning two limits to a certain parameter, and in asserting that, with some specified probability, the true value of the parameter will be situated between these limits, which constitute the confidence interval.

The bootstrap is a powerful technique for assessing the accuracy of a parameter estimator in situations where conventional techniques are not valid. The bootstrap does with a computer what the experimenter would do in practice if it were possible: he or she would repeat the experiment. With the bootstrap, the observations are randomly reassigned and estimates recomputed. Given n estimates of the parameter we compute the empirical distribution function assigning probabilities 1/n, 2/n.....n-1/n to the ordered values of the estimates. We then generate n uniform random numbers on (0, 1). Assign the k^{th} order statistic as the sample value if $u_i \leq$ k/n. The recomputations are done a thousand times and treated as repeated experiments.

III. THE FINITE SAMPLE SIMULATION

3.1 Influence of the Source Distributions

From the experiment of Bach and Jordan, 18 density functions were generated. These are:

(a) Student's T with 3 degrees of freedom;
(b) double exponential;
(c) uniform;
(d) Student's T with 5 degrees of freedom;
(e) exponential;
(f) mixture of two double exponentials;
(g)-(h)-(i) symmetric mixtures of two Gaussians: multimodal, transitional and unimodal;
(j)-(k)-(l) nonsymmetric 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.

The performance of Kernel ICA is robust with respect to the source distributions. Performance is similar across multimodal (f, g, j, m, p), unimodal (a, b, d, e, i, l, o, r) and transitional (c, h, k, n, q) distributions. The Kernel ICA algorithms are particularly insensitive to asymmetry of the pdf when compared to the other algorithms. Since the performance is similar across multimodal, unimodal and transitional distributions, and that the Kernel ICA is insensitive to asymmetry of the pdf, this study attempts to utilize 6 distributions only, some unimodal and some symmetric. The 6 pdf's are (a) Student with 3 degrees of freedom; (b) double exponential; (d) Student with 5 degrees of freedom; (e) exponential; (i) symmetric mixtures of two Gaussians; and (o) symmetric mixtures of four Gaussians;

3.2 Kernel-ICA Algorithm

The experiments made use of the same basic procedure for generating data:

(1) N samples of each of the M sources were generated according to their pdf's and placed into an $M \times N$ matrix **x**.
(2) A random mixing matrix **A** was chosen, with random but bounded condition number
(3) A matrix $\tilde{\mathbf{y}}$ of dimension $M \times N$ was formed as the mixture $\tilde{\mathbf{y}} = \mathrm{Ax}$.

The Kernel ICA outputs a demixing matrix **W** which can be applied to the matrix **Y** to recover estimates of the independent components. A high-level description of the Kernel ICA algorithm for estimating the parameter matrix **W** in the ICA model is presented below.

Algorithm Kernel ICA-KCCA

Input: Data vectors $y_1, y_2 \ldots . y_N$
Kernel $K(x,y)$

1. Whiten the data.
2. Minimize (with respect to W) the contrast function $C(\mathbf{W})$ defined as:

a. Compute the centered Gram matrices $K_1, K_2 \ldots K_n$ of the estimated sources $\{x_1, x_2 \ldots . x_N\}$, where $x_i = W y_i$

b. Define $\lambda_F^K(K_1, K_2 \ldots K_M)$ as the minimal eigenvalue of the generalized eigenvalue equation.

c. Define $C(W) = \lambda_F^K(K_1, K_2 \ldots K_M) = -\frac{1}{2} \log \lambda_F^K(K_1, K_2 \ldots K_M)$

Thus, given the set of vectors $y_1, y_2 \ldots . y_N$, and a given initial parameter **W**, we set $x_i = \mathbf{W}y_i$, for each i and thereby form a set of estimated source vectors $\{x_1, x_2 \ldots . x_N\}$. The M components of these vectors yield a set of M Gram matrices $K_1, K_2, \ldots . K_M$, and these Gram matrices (which depend on **W**) define the contrast function C(**W**). The ICA algorithm was obtained by minimizing this function with respect to **W**.

In the first series of experiments, we tested Kernel ICA on a two-component ICA problem, with all 6 unimodal source distributions. In particular, the two source distributions for the ICA model are the same. For each of the 6 sources (a, b, d, e, i, o) 20, 30, 60 and 90 sample sizes were generated, resulting to 30 values of the contrast function C(**W**) each sample size. Then 1000 bootstrap replicates were generated utilizing these 30 C(**W**) values.

In the next series of experiments, we tested the algorithms with 2, 4 and 6 components whose source distributions were chosen at random from the 6 unimodal and symmetric sources.

IV. RESULTS AND DISCUSSION

This study utilizes 6 distributions only, all unimodal. The 6 pdf's are (a) Student with 3 degrees of freedom; (b) double exponential; (d) Student with 5 degrees of freedom; (e) exponential; (g-h-i) symmetric mixtures of two Gaussians; and (m-n-o) symmetric mixtures of four Gaussians.

4.1 Experiment 1 (Two-Component ICA with Same Source PDF's)

In the first experiment, two-component signals of length 250 were generated for signals coming from the same distribution, i.e. two student's t with 3 degrees of freedom (aa), two double exponentials (bb), two student's t with 5 degrees of freedom (dd), two exponentials (ee), two symmetric mixtures of two Gaussians (ii) and two symmetric mixtures of four Gaussians etc.

First, we generated two-component signal aa. So we have a 2 *x* 250 original signals which were mixed using a random mixing matrix of size 2 *x* 2. These mixed signals were then fed into the Kernel ICA program which resulted to one value of the contrast function C(**W**). Then the same procedure was applied twenty times producing twenty values of C(**W**). After this, bootstrap procedure was applied to the 20 values of the contrast function where one thousand (1000) bootstrap replicates were utilized for estimating the true mean of C(**W**). Then the experiment continued utilizing signals of sizes 2 *x* 30, 2 *x* 60, and 2 *x* 90. Table 1 below shows the results of the first experiment.

Table 1. The mean and variance of the values of C(W) for 1000-bootstrapped Two-component ICA (same pdf's)

PDF	MEAN				VARIANCE			
	n = 20	n = 30	n = 60	n=90	n = 20	n = 30	n = 60	n=90
aa	0.3126538	0.2361176	0.1679274	0.1354645	0.00010202	0.00005219	0.00002856	0.00000880
bb	0.3262286	0.2584961	0.1802871	0.1361412	0.00015163	0.00005190	0.00004089	0.00001526
dd	0.3210324	0.2523636	0.1676068	0.1312062	0.00010092	0.00004898	0.00001803	0.00000823
ee	0.3106494	0.261729	0.1601975	0.1369996	0.00006473	0.00009152	0.00001960	0.00005031
ii	0.2971559	0.2302171	0.1782419	0.1371616	0.00006926	0.00004657	0.00003805	0.00001894
oo	0.2992455	0.2501374	0.1812705	0.1532209	0.00007458	0.00007381	0.00004777	0.00003318

It can be seen that the mean tends to decrease as the number of sample of the contrast function increases. For example, for the two-component ICA aa (two student's t with 5 degrees of freedom), the means of the contrast function C(W) are 0.3126538, 0.2361176, 0.1679274, and 0.1354645 for a 20, 30, 60 and 90 sample sizes respectively. The same situation is true to the remaining two-component (the same pdf) ICA. **Finite sample calculations show that the means systematically *decrease to zero as expected*, i.e. that the F-correlation tends to zero showing that as n increases, the *contrast function is able to separate independent components more effectively.***

The variance of C(W) also decreases as we increase the sample size from 20 to 90 samples. For the signals aa, the variances are 0.00010202, 0.00005219, 0.00002856 and 0.00000880 for the sample sizes of 20, 30, 60, and 90 respectively. Again, one can notice that for the signals bb, dd, ee, ii, oo, the variance of C(W) all decreases as the sample size increases from 20, to 60 and to 90.

4.2 Experiment 2 (Two-Component ICA with Different Source PDF's)

The second experiment was similar to the first experiment, only that it utilized data coming from two different distributions. Thus, twenty (20), thirty (30), sixty (60) and ninety (90) two-component signals were generated for signals coming from two different distributions (i.e. one student's-t and one double exponential (ab), one student's- t and one exponential (ae) etc.). These were again mixed and fed into the Kernel ICA program producing 30 of the contrast function C(W). Again, bootstrap procedure was applied to the output of the Kernel ICA.

For easy reference, we will summarize the following pairs and its corresponding densities:

ab - one student's t (d.f. = 3) and one double exponential
ad - two student's t with (d.f. = 3) and (d.f. = 5)
ae - one student's t (d.f. = 3) and one exponential
ai - one student's t (d.f. = 3) and one mixture of 2 gaussians
ao - one student's t (d.f. = 3) and one mixture of 4 gaussians
bd - one double exponential and one student's t (d.f. = 5)
be - one double exponential and one exponential
bi - one double exponential and one mixture of 2 gaussians
bo - one double exponential and one mixture of 4 gaussians
de - one student's t (d.f. = 5) and one exponential
di - one student's t (d.f. = 5) and one mixture of 2 gaussians
do - one student's t (d.f. = 5) and one mixture of 4 gaussians

ei - one exponential and one mixture of 2 gaussians
eo - one exponential and one mixture of 4 gaussians
io - one mixture of 2 gaussians and one mixture of 4 gaussians

Table 2 shows the results of the second experiment.

Table 2. The mean and variance of the values of C(W) for 1000-bootstrapped Two-component ICA's (different pdf's)

PDF	MEAN				VARIANCE			
	n = 20	n = 30	n = 60	n=90	n = 20	n = 30	n = 60	n=90
ab	0.311922	0.24932	0.172755	0.147744	0.00009535	0.00004486	0.00003364	0.00002671
ad	0.32006	0.242398	0.174149	0.138707	0.00010962	0.00006387	0.00002333	0.00001264
ae	0.323872	0.253826	0.172791	0.129802	0.00013733	0.00010406	0.00003663	0.00002033
ai	0.315721	0.252701	0.174118	0.133652	0.00009433	0.00004987	0.00001567	0.00002286
ao	0.32375	0.254112	0.172079	0.139584	0.00015364	0.00004883	0.00004491	0.00001719
bd	0.327904	0.250709	0.176238	0.13913	0.00014544	0.00004097	0.00002756	0.00001782
be	0.316789	0.254166	0.172532	0.131175	0.00010817	0.00007060	0.00004752	0.00002067
bi	0.336079	0.239053	0.173407	0.142518	0.00011617	0.00005030	0.00003408	0.00002656
bo	0.333192	0.241044	0.176988	0.141315	0.00017667	0.00005867	0.00005744	0.00001786
de	0.318091	0.257829	0.175305	0.131627	0.00005959	0.00008570	0.00003636	0.00001600
di	0.30341	0.25112	0.16532	0.136707	0.00010109	0.00004791	0.00001760	0.00001478
do	0.306477	0.258409	0.171458	0.143924	0.00010508	0.00004821	0.00002942	0.00002426
ei	0.314765	0.249945	0.180038	0.144543	0.00012855	0.00004257	0.00004414	0.00004502
eo	0.310241	0.239844	0.169062	0.134372	0.00006442	0.00003559	0.00002196	0.00002229
io	0.312499	0.260838	0.177868	0.144687	0.00012723	0.00006087	0.00003924	0.00002669

It can also be observed, similar to experiment 1, that the mean and variance of C(W) tends to decrease as the number of sample of the contrast function increases.

4.3 Experiment 3 (Four-Component ICA with Different Source PDF's)

The third experiment utilized data coming from four different distributions. Twenty (20), thirty (30), sixty (60) and ninety (90) signals were generated for signals coming from four different distributions (i.e. abde corresponds to one student's t with 3 degrees of freedom, one double exponential , one student's t with 5 degrees of freedom and one exponential, etc.). Again, these four-component signals were fed into the Kernel ICA program producing 30 values of the contrast function C(W). Then bootstrap procedure was applied to these 30 contrast functions.

Table 3 below shows the results of the third experiment. Also included in table is the result of the six-component input signal result.

Table 3. The mean and variance of the values of C(W) for 1000-bootstrapped Four-component ICA (different pdf's) and six-component ICA

PDF	MEAN				VARIANCE			
	n = 20	n = 30	n = 60	n=90	n = 20	n = 30	n = 60	n=90
Abde	0.564141	0.458337	0.312346	0.253701	0.00005980	0.00004501	0.00001980	0.00001729
Abei	0.562385	0.450788	0.319879	0.255989	0.00004552	0.00004214	0.00002731	0.00001631
Abdi	0.566942	0.452426	0.306974	0.244067	0.00004840	0.00004154	0.00002525	0.00001257
Abdo	0.557911	0.448588	0.302056	0.251451	0.00005133	0.00005956	0.00002160	0.00001606
Adei	0.552133	0.451729	0.306142	0.260737	0.00004055	0.00006266	0.00002574	0.00002366
Adeo	0.564709	0.445342	0.319126	0.263393	0.00005335	0.00003910	0.00003469	0.00002023
Adio	0.557681	0.449652	0.303747	0.248956	0.00006407	0.00002516	0.00001339	0.00001529
Abdeio	0.739197	0.609995	0.422278	0.345686	0.00001922	0.00002304	0.00000989	0.00001098

Similar with the two experiments, the mean of C(**W**) tends to decrease as the number of sample of the contrast function increases. This is to be expected since the objective is to put the contrast function C(W) close to zero. The variance of C(**W**) also decreases as we increase the sample size.

4.4 Relationship of the Variance of the Contrast Function F-Correlation to the Number of Sample

One objective of this paper is to find the relationship between the variance of the contrast function versus the number of samples. It is noted at once that the relationship between the variances and the sample sizes is non-linear i.e. that the variance does not decrease as a linear function of n. Thus, a **non-linear regression model of the form**:

$$\text{Variance} = e^{\text{intercept}} (\text{no. of sample})^{\alpha} \text{ or}$$
$$\text{Variance} = A n^{\alpha}$$

between the variance of the C(W) and the number of samples is tried. The results are shown in Table 4. Observe that the alpha's are all raised to negative powers. We have observed the same situation with the all the other two-component ICA's as well as the four-component ICA. Thus, for a two-component ICA with generally different pdf's, we can conclude that if the number of sample is very large, the variance of C(**W**) converges to zero.

Table 4. The coefficients of the regression model Variance = $e^{\text{intercept}}$ (no. of sample)$^{\alpha}$

PDF	Intercept	A
aa	-2.03318	-1.49615
bb	-2.17958	-1.31948
dd	-1.88527	-1.62714
ee	-3.44953	-0.52766
ii	-3.16799	-0.76211
oo	-3.36291	-0.55412
ab	-3.09960	-0.77058
ad	-2.07890	-1.43913
ae	-2.10867	-1.31146
ai	-2.70198	-1.06987
ao	-2.31986	-1.22228
bd	-2.34640	-1.25152
be	-2.64113	-1.00765
bi	-2.84188	-0.90762
bo	-2.18906	-1.26212
de	-2.85544	-0.93776
di	-2.34679	-1.30750
do	-2.84886	-0.92870
ei	-3.31929	-0.56623
eo	-3.34158	-0.70403
io	-2.69818	-0.96739
abde	-3.06350	-0.88803
abei	-3.42254	-0.67247
abdi	-3.14062	-0.86515
abdo	-3.03927	-0.89354
adei	-3.60576	-0.52029
adeo	-3.54385	-0.56202
adio	-3.08203	-0.94513
abdeio	-3.98188	-0.52076

4.5 Relationship of the Average Amari Error and the Number of Sample

The "Amari Error" is a measure of performance of an ICA algorithm. A value of zero indicates perfect match between the found solution and the sources. The greater the value of the this measure, the poorer the performance. Table 5 below shows the relationship between the average Amari error and the number of sample.

Table 5. The Average Amari Errors at different sample sizes

PDF	AMARI ERRORS				
	n = 20	n = 30	n = 60	n=90	AVERAGE
aa	0.00089000	0.00204000	0.00148000	0.00189667	**0.001577**
bb	0.00094000	0.00138333	0.00087333	0.00099333	**0.001047**
dd	0.00145333	0.00152667	0.00060667	0.00111667	**0.001176**
ee	0.00153333	0.00108667	0.00099000	0.00100333	**0.001153**
ii	0.00230333	0.00123000	0.00123333	0.00117333	**0.001485**
oo	0.00058000	0.00092667	0.00155000	0.00067333	**0.000933**
ab	0.00098440	0.00135670	0.00115330	0.00069670	**0.001048**
ad	0.00108860	0.00122000	0.00055000	0.00134000	**0.001050**
ae	0.00199670	0.00139000	0.00094670	0.00075330	**0.001272**
ai	0.00152500	0.00141670	0.00100670	0.00149330	**0.001360**
ao	0.00184670	0.00156330	0.00108330	0.00115330	**0.001412**
bd	0.00090000	0.00181000	0.00066670	0.00112330	**0.001125**
bè	0.00120670	0.00189670	0.00070000	0.00150670	**0.001328**
bi	0.00250330	0.00157330	0.00071000	0.00135670	**0.001536**
bo	0.03280140	0.00129330	0.00044330	0.00119330	**0.008933**
de	0.00102800	0.00112330	0.00158670	0.00081670	**0.001139**
di	0.00136110	0.00147670	0.00053670	0.00199670	**0.001343**
do	0.00140670	0.00163000	0.00155330	0.00131000	**0.001475**
ei	0.00107000	0.00104330	0.00089670	0.00122330	**0.001058**
eo	0.00159670	0.00105000	0.00064000	0.00149670	**0.001196**
io	0.00141670	0.00117000	0.00093330	0.00225330	**0.001443**
abde	0.00647333	0.00586000	0.00636000	0.00572330	**0.006104**
abei	0.00576333	0.00476667	0.00475667	0.00507000	**0.005089**
abdi	0.00695333	0.00537000	0.00496000	0.00494000	**0.005556**
abdo	0.00525333	0.00537000	0.00609000	0.00578670	**0.005625**
adei	0.00534667	0.00538000	0.00475000	0.00479000	**0.005067**
adeo	0.00542333	0.00621667	0.00579000	0.00570330	**0.005783**
adio	0.00551667	0.00651000	0.00625333	0.00603330	**0.006078**
abdeio	0.00461000	0.00612000	0.00557333	0.00509330	**0.005349**

Looking at the rightmost column of the above tables, we can see that the average Amari errors are generally of the order 10^{-3}.

4.6 Percentile Points of the F-Correlation Contrast Function for Small Sample Sizes

The following table gives the small sample distribution of the contrast function C(W) based on the bootstrap simulation experiment. The table can be used for testing hypothesis about the values of C(W) contrast function for small sample sizes.

Table 6: Percentile Points of the Contrast Function C(W)

	Two-component ICA (Same Pdf's)							
	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99
n=20	0.288733	0.292433	0.295333	0.298817	0.323383	0.326883	0.32945	0.332683
n=30	0.230500	0.233117	0.23595	0.2384	0.25825	0.261167	0.26385	0.266483
n=60	0.160433	0.16205	0.1637	0.165633	0.179867	0.182133	0.1843	0.18625
n=90	0.128650	0.129867	0.131167	0.132667	0.144217	0.145967	0.147533	0.1492
	Two-component ICA (Different Pdf's)							
	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99
n=20	0.29434	0.29804	0.301133	0.304807	0.331947	0.3359	0.339793	0.343647
n=30	0.23360	0.23602	0.23836	0.24088	0.2599	0.262913	0.2655	0.268767
n=60	0.16106	0.16268	0.164387	0.166307	0.180967	0.183187	0.185267	0.187567
n=90	0.12846	0.129847	0.131173	0.132713	0.144613	0.146413	0.148067	0.150207
	Four-component ICA (Different Pdf's)							
	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99
n=20	0.543943	0.546729	0.549014	0.551757	0.569971	0.572514	0.575086	0.577529
n=30	0.436086	0.438257	0.440314	0.442643	0.459557	0.462414	0.464643	0.467114
n=60	0.298786	0.300614	0.302143	0.3038	0.316271	0.317971	0.319643	0.3215
n=90	0.244700	0.246057	0.247343	0.2488	0.259429	0.260943	0.262343	0.264071
	Six-component ICA (Different Pdf's)							
	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99
n=20	0.7296	0.7308	0.7321	0.7335	0.7453	0.7467	0.7479	0.7495
n=30	0.5985	0.6005	0.6017	0.6039	0.616	0.6176	0.6188	0.6209
n=60	0.4205	0.4243	0.4265	0.4244	0.4264	0.4248	0.4273	0.4229
n=90	0.3379	0.3391	0.3406	0.3414	0.3501	0.3513	0.3518	0.3533

Thus, if one wishes to test the hypothesis that C(W) =0 at the .05 level of significance with n = 90 or more, knowing that there are at least four components in the signal, then the hypothesis is accepted whenever the computed correlation is less than or equal to 0.260943, and conclude that the signals are independent.

V. CONCLUSION

On the basis of the foregoing analyses, the following conclusions are warranted:

1. The variance of the computed C(W) function decreases as $O(n^{\alpha})$, (in n^{α}, α is a negative number ranging from – 1.68 to – 0.52 for all samples of size n). This means that increasing the sample size will decrease the variance of the contrast function C(W). There are, therefore, indications that the contrast function C(W) converges in mean square to a fixed value for large n.

2. Since the variances computed for signals coming from different pdf's are generally smaller than the variances computed for signals coming from the same pdf's, the contrast function C(W) is able to separate signals from different source distribution functions more efficiently than when the signals come from the same source distribution functions.

3. The small sample sampling distribution of the contrast function C(**W**) values appear to range within a very narrow band and varies with different sample sizes.

4. Since the *normalized* Amari errors are generally of the order 10^{-3}, then the contrast function C(**W**) is able to separate signals with 99.9% accuracy. (Note that the non-normalized Amari error can range from 0 to (m-1) but the *normalized* Amari error ranges from 0 to 1).

Acknowledgement

We wish to thank the editors of the Philippine Statistician and the anonymous referee for their helpful comments and suggestions.

References

BACH, Francis and JORDAN, Michael. 2002. Kernel Independent Component Analysis, Journal of Machine Learning Research 3, pp. 1-48.

DAVENPORT, Mark. 2005. Kernel Methods and Their Application to PCA and ICA.

HYVARINEN, Aapo and OJA, Erkki. 1997. A Fast Fixed-Point Algorithm for Independent Component Analysis, Neural Computation, 9:1483-1492.

JORDAN, Michael. 2004. Stat241B: Advanced Topics in Learning and Decision Making - Properties of Kernels and the Gaussian Kernel, Lecture of 2/04/2004.

JORDAN, Michael. 2004. Stat241B: Advanced Topics in Learning and Decision Making – Multiple Kernels and Reproducing Kernel Hilbert Spaces.

XU, Jian-Wu, ERDOGMUS, Deniz, JENSSEN, Robert and PRINCIPE, Jose. 2004. An Information-Theoretic Perspective to Kernel Independent Component Analysis