A STATISTICAL APPROACH TO TESTING MUTUAL INDEPENDENCE OF ICA RECOVERED SOURCES

Kai-Chun Chiu, Zhi-Yong Liu and Lei Xu

Department of Computer Science and Engineering, The Chinese University of Hong Kong
Shatin, New Territories, Hong Kong, P. R. China
{kcchiu,zyliu,lxu}@cse.cuhk.edu.hk

ABSTRACT

When independent component analysis (ICA) is applied on real data, the source signals as well as the mixing matrix are blind to users. In such case testing for mutual independence of estimated source signals is of vital importance. In this paper, we illustrate and discuss how testing mutual independence of estimated sources signals can be done in the context of testing multivariate uniformity.

1. INTRODUCTION

Independent component analysis (ICA) [1] has been adopted for separating blind signals in a variety of tasks such as radio-communication, speech enhancement, seismic signals, nuclear reactor monitoring and airport surveillance.

Mathematically, ICA can be expressed as

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

where $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$ denotes the vector of observed random variables, $\mathbf{s} = (s_1, s_2, \cdots, s_n)^T$ is the vector of the independent blind source signals, and $\mathbf{A} \in \mathbb{R}^{n \times k} (n \geq k)$, is an unknown non-singular mixing matrix.

Without knowing the source signals $\mathbf{s}$ and the mixing matrix $\mathbf{A}$, the task of ICA is to find a so-called demixing matrix $\mathbf{W}$ to recover the original signals from the observations $\mathbf{x}$ by the following linear transform:

$$\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{s} = \mathbf{V}\mathbf{s}, \quad \mathbf{V} = \mathbf{W}\mathbf{A} \tag{2}$$

where $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ denotes the estimated source signals such that either $\mathbf{y}$=$\mathbf{s}$ or $\mathbf{y}$ recovers $\mathbf{s}$ only up to constant unknown scales and any permutation of indices.

The work described in this paper was fully supported by a grant from the Research Grant Council of the Hong Kong SAR (Project No: CUHK 4169/00E).

However, independence of the estimated source signals as defined by

$$P_{Y_1 \cdots Y_n}(y_1, \cdots, y_n) = P_{Y_1}(y_1) \cdots P_{Y_n}(y_n). \tag{3}$$

cannot be verified in real practice because the mixing matrix $\mathbf{A}$ as well as the source signals $\mathbf{y}$ are all blind.

Moreover, we cannot directly evaluate the independence of estimated source signals via

$$KL(\mathbf{W}) = \int p(\mathbf{y}) \log \frac{p(\mathbf{y})}{\prod_{i=1}^{k} p(y_i)} d\mathbf{y} \tag{4}$$

because the joint distribution of sources cannot be estimated.

Literature concerning test for independence was not common. Test statistic based on comparing the residual of the empirical characteristic function and true characteristic function was suggested by [2]. It is based on the property of the Fourier-Stieltjes transform that one characteristic function corresponds to one distribution. Therefore, closeness of two distributions can be restated in terms of closeness of their characteristic functions. However, computational cost increases drastically as the number of sources increases. On the other hand, although the test statistics based on the generalized least squares criterion proposed by [3] was supported by some empirical evidence, further theoretical results are needed on the distribution and convergence properties.

In view of the lack of appropriate test for source independence, a new approach based on testing multivariate uniformity of suitably transformed estimated source signals is proposed in this paper. The rest of the paper is organized in the following way. Section 2 establishes the theoretical equivalence between testing for independence of the source signals and uniformity of the transformed signals. Section 3 outlines test statistics and test procedures. Section 4 discusses several issues related to using the uniformity test for independence. Section 5 is devoted to experimental simulation and section 6 concludes the paper.

2. TESTING MULTIVARIATE UNIFORMITY AS AN ALTERNATIVE FOR INDEPENDENCE

By the famous integral transformation theorem [4], we know that if Y is a real-valued random variable with continuous cumulative distribution function F, then $Z = F(Y)$ has a uniform distribution on the interval $[0, 1]$, i.e., Z is a $U[0, 1]$ random variable.

Moreover, if we consider the estimated sources are mutually independent and each of them if transformed by its corresponding cumulative distribution function, then all source signals would have multivariate uniform distribution in the multidimensional unit cube $[0, 1]^k$. Conversely, if any two of the estimated sources are not independent, then the joint density of the transformed sources would be [5]

$$\begin{aligned} & P_{Z_1 \cdots Z_n}(z_1, \cdots, z_k) \\ = \; & P_{Z_1 \cdots Z_n}(F_{Y_1}(y_1), \cdots, F_{Y_k}(y_k)) \\ = \; & \frac{P_{Y_1 \cdots Y_n}(y_1, \cdots, y_k)}{\prod_{i=1}^{k} P_{Y_i}(y_i)} \neq 1. \end{aligned} \tag{5}$$

It implies that the joint distribution of the transformed signals would not be multivariate uniform. Thus, testing multivariate uniformity of transformed signals can be used as an alternative for testing for independence of the estimated signals.

3. THE TEST FOR MULTIVARIATE UNIFORMITY

Early studies on testing uniformity in the unit interval $[0, 1]$ were referred to [6, 7]. [4] gave a comprehensive review. However, testing uniformity of random samples in the multidimensional unit cube $\bar{C}^k (k \geq 2)$, where

$$\begin{aligned} \bar{C}^k = [0, 1]^k = \; & \{\mathbf{x} = (x_1, \cdots x_k)^T \in \mathbb{R}^k : \\ & 0 \leq x_i \leq 1, i = 1, \cdots k\}, \end{aligned}$$

seems to have received less attention in the literature of statistics. Recent development [5] in the area has indeed provided an efficient tool that enables us to perform test related to independence.

3.1. The Null and Alternate Hypotheses

The null hypothesis for testing the uniformity of random samples $\mathcal{P} = z_1, \cdots, z_n \subset \bar{C}^d$ where can be stated as $H_0 : z_1, \cdots, z_n$ are uniformly distributed in the unit cube $\bar{C}^k$. The alternative hypothesis H_1 implies rejection of H_0. A test for multivariate uniformity can be performed by determining whether the value of a test statistic is unlikely under the null hypothesis.

3.2. Properties of the Test Statistic

The test statistic is derived from the generalized $\mathcal{L}^2$ type discrepancy $D(\mathcal{P}^2)$ [8]. Under the null hypothesis, the statistic

$$A_n = \sqrt{n} \left[(U_1 - M^k) + 2(U_2 - M^k) \right] / (5\sqrt{\zeta_1}) \tag{6}$$

is shown in [5] to converge to the standard normal distribution $\mathcal{N}(0, 1)$ as $n \to \infty$, where U_1 given by

$$U_1 = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{k} \left[M + \beta^2 \mu(z_{ij}) \right], \tag{7}$$

converges almost surely to M^k by the strong law of large numbers.
U_2 is a second-order U-statistic given by

$$\begin{aligned} U_2 \; = \; & \frac{2}{n(n-1)} \sum_{i<l}^{n} \prod_{j=1}^{k} (M + \beta^2 [\mu(z_{ij}) + \mu(z_{lj}) \\ & + \frac{1}{2} B_2(|z_{ij} - z_{lj}|) + B_1(z_{ij}) B_1(z_{lj})]) \end{aligned} \tag{8}$$

for $i, l = 1, \cdots, n$, and converges to M^k by the strong law of large numbers for general U-statistics.
ζ_1 is given by

$$\zeta_1 = (M^2 + \beta^4 c^2)^k - M^{2k} \quad \text{where } c^2 = \int_0^1 \mu(x)^2 dx \tag{9}$$

The $B_1(\cdot)$ and $B_2(\cdot)$ in (8) are the first and second degree Bernoulli polynomials, respectively:
$B_1(x) = x - \frac{1}{2}$ and $B_2(x) = x^2 - x + \frac{1}{6}$.

Three specific discrepancies, namely symmetric, centered and star discrepancy, can be obtained by taking different choices of the constants β,M and the function $\mu(x)$. We favor the symmetric discrepancy because it is more powerful than its counterparts. Since the power of a test is equal to $(1 - \beta)$ where β is the Type II error, this implies that the probability for the test statistic with symmetric discrepancy to make a mistake when the sample actually comes from a non-uniform distribution is lower. Consequently, we have

$$\mu(x) = -\tfrac{1}{2}(x^2 - x + \tfrac{1}{6}), \quad \beta = 2, \quad M = \tfrac{4}{3}.$$

Larger values of the statistic $|A_n|$ imply rejection for H_0. Critical values of $|A_n|$ at $\alpha = 5\%$ for the asymptotic as well as finite-sample distributions are show in Table 1(A) [5].

4. SPECIFIC ISSUES RELATED TO THE UNIFORMITY TEST FOR INDEPENDENCE

4.1. Transforming Source Signals by CDF

To enable the uniformity test to be used, the each source should be transformed by its respective cumulative distribution function (cdf). Since the actual cdf has to be estimated, we would use the empirical cdf which is an unbiased estimator of the true cdf. Suppose that $\{y_1, \cdots, y_n\}$ is a dataset of observed values from a real-valued random variable, source transformation can be achieved by means of the empirical cdf defined by

$$F(y) = \frac{\#\{i\{1, 2, \cdots, n\} : y_i \leq y\}}{n} \quad \text{for } y \in \mathbb{R}. \quad (10)$$

Thus, $F(y)$ gives the fraction of values in the dataset less than or equal to y.

4.2. Critical Values by Monte Carlo Simulation

Although testing for independence can be alternatively done via testing multivariate uniformity, two issues still needs to be resolved before the test can be used for testing for source independence for ICA.

The first one is concerned with the effect of the transformation by cdf on the critical values of the original test statistic. This is motivated by the empirical fact shown in Table 2. We observe that if the original critical values 1.96 or -1.96 for testing bivariate uniformity as shown in Column 2 & 3 of Table 1(A) were used, then the computed A_n would not reject independence even if the sources are significantly dependent.

Figure 1 compares the differences between testing multivariate uniformity and testing for independence. As shown by the connection between the two grey boxes, the inappropriateness of using the original critical values for testing for independence may be due to the univariate transformation by cdf.

Since the critical values shown in Table 1(A) is not suitable for testing for independence, we have to determine new critical values. This can be done by Monte Carlo simulation. For each sample size of 25,50,100,200 and 1000, we generate 1000 samples to obtain a new set of critical values. Results are summarized in Table 1(B).

By comparing Table 1(A) & (B), we can see that at $\alpha = 5\%$ and the critical values are very different. The smaller critical values of Table 1(B) reveals the effect of source transformation by their respective cdf. Figure 2 shows two typical independent sources before and after transformation by their respective cdf.

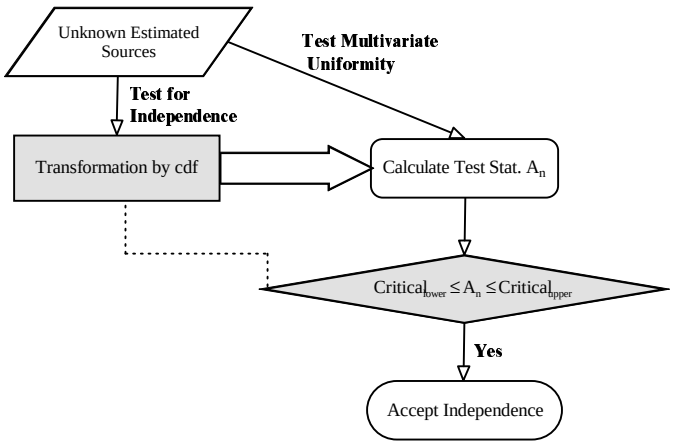
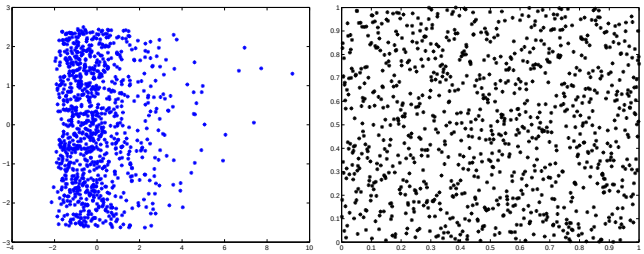


Figure 1: A comparative view of the approach for testing uniformity and that modified for testing for independence



(a) Before transformation (b) After transformation

Figure 2: 2-dimensional view of independent sources before and after uniform transformation

4.3. Relationship between Dependence and Non-Uniformity

The second issue concerns whether greater dependence always results in greater distortion on multivariate uniformity, i.e., there is a monotonic relationship between the dependence and non-uniformity. When sources are not independent, it is known that the transformed sources would deviate from multivariate uniformity. To determine whether greater dependence results in greater distortion on the test statistic for multivariate uniformity, we consider a sample consisting of 1000 2-dimensional uniformly distributed data. Correlation is introduced via a rotation of two independent sources. We examine the effect of 2nd, 3rd and 4th order dependence on the test statistic A_n for bivariate uniformity. During simulation, we alter the degree of second order correlation as well as higher order dependence via the rotation matrix. Typical results obtained are shown in Table 2. As shown in Table 2, a reasonable pattern observed is higher dependence results in greater distor-

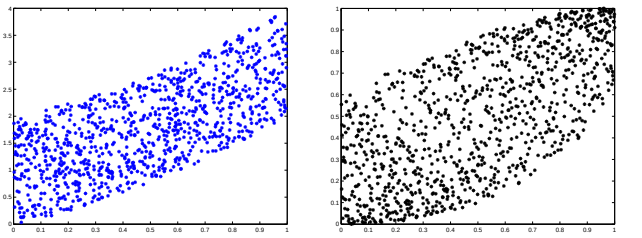
Table 1: Comparison of critical values for testing multivariate uniformity and those for testing for independence at $\alpha = 5\%$.

(A) Critical values for testing multivariate uniformity						
Sample	k=2		k=5		k=10	
Size	Upper	Lower	Upper	Lower	Upper	Lower
25	2.1836	-1.9566	2.1698	-1.9750	2.3549	-1.9266
50	2.0585	-1.9969	2.0814	-1.9143	2.0965	-1.8729
100	2.0078	-1.9102	2.0271	-1.9372	2.0861	-1.8981
200	1.9897	-1.9279	2.0671	-1.9199	2.0855	-1.9743
∞	1.9600	-1.9600	1.9600	-1.9600	1.9600	-1.9600
(B) Critical values for testing for independence						
25	-0.3057	-0.6797	-0.4334	-1.1184	-0.4702	-1.5962
50	-0.1958	-0.5215	-0.1589	-0.9189	-0.3258	-1.1894
100	-0.1089	-0.3658	-0.0900	-0.6319	-0.0110	-0.9506
200	-0.0408	-0.2943	0.0136	-0.4991	0.0398	-0.8472
1000	0.0631	-0.1743	0.0568	-0.3770	0.1851	-0.3760

Table 2: Results by simulation to examine effect of dependence on distortions of the test statistic for bivariate uniformity

Correlation Coefficient						
2nd order	3rd order		4th order			$\lvert A_n \rvert$
0.0098	0.0199	0.0220	0.0267	0.0365	0.0275	0.0233
0.0906	0.0775	0.0781	0.0658	0.0612	0.0757	0.2463
0.1883	0.1724	0.1765	0.1558	0.1583	0.1752	0.2931
0.2806	0.2621	0.2698	0.2411	0.2511	0.2674	0.4531
0.3656	0.3449	0.3558	0.3199	0.3373	0.3503	0.6777
0.4423	0.4196	0.4331	0.3909	0.4154	0.4233	1.1856
0.5103	0.4858	0.5014	0.4540	0.4848	0.4869	1.7571
0.5698	0.5438	0.5609	0.5093	0.5457	0.5417	2.3568
0.6213	0.5941	0.6124	0.5573	0.5986	0.5888	2.9688
0.7040	0.6750	0.6944	0.6345	0.6837	0.6638	4.1518

tion on bivariate uniformity and hence greater value of A_n. Figure 3 shows two dependent sources before and after the transformation by their respective cdf and Figure 4 shows results of two dependent but uncorrelated sources. Although Figure 4(a) [obtained from de-correlation of Figure 3(a)] after transformation still cannot achieve bivariate uniformity, it is more close to bivariate uniformity as compared with Figure 3(b) due to de-correlation.



(a) Before transformation (b) After PCA transformation

Figure 3: 2-dimensional view of dependent sources before and after uniform transformation

5. SIMULATIONS

Suppose three samples of 5-dimensional vectors representing signals of 2 sub-Gaussian sources and 3 super-Gaussian sources are estimated via a typical ICA algorithm. The two sub-Gaussian sources are generated with uniform distribution while the three super-Gaussian sources consist of one lognormal and two speech signals [1]. The simulation is done with sample size $n = 50, 200$ and 1000 respectively. Independence is first verified via checking if every row of **WA** has one and only one sig-

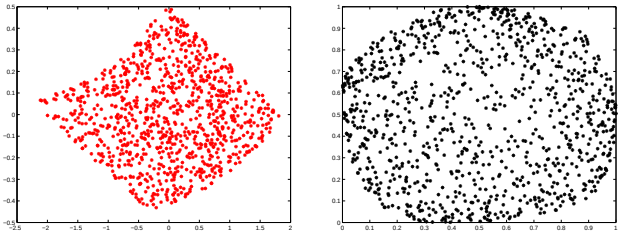
[1]Source data are available for download at http://sweat.cs.unm.edu/~bap/domos.html

Table 3: Computed test statistic and statistical inference for independent sources.

Sample	95% Computed A_n		Critical A_n at $\alpha = 5\%$		Statistical	Empirical
Size	Upper	Lower	Upper	Lower	Inference	$\hat{\alpha}/\hat{\beta}$
50	-0.2057	-0.6128	-0.1589	-0.9189	Reject H_1	$\hat{\alpha} = 0\%$
200	-0.0147	-0.5551	0.0136	-0.4991	Reject H_1	$\hat{\alpha} = 2\%$
1000	0.1174	-0.4123	0.0568	-0.3770	Reject H_1	$\hat{\alpha} = 3\%$

Table 4: Computed test statistic and statistical inference for dependent sources.

Sample	95% A_n Outside		Critical A_n at $\alpha = 5\%$		Statistical	Empirical
Size	Upper	Lower	Upper	Lower	Inference	$\hat{\alpha}/\hat{\beta}$
50	-0.1803	-0.8121	-0.1589	-0.9189	Reject H_0	$\hat{\beta} = 5\%$
200	-0.0984	-0.5319	0.0136	-0.4991	Reject H_0	$\hat{\beta} = 3\%$
1000	0.0436	-0.2943	0.0568	-0.3770	Reject H_0	$\hat{\beta} = 6\%$



(a) Before transformation (b) After transformation

Figure 4: 2-dimensional view of uncorrelated dependent sources before and after uniform transformation

nificant non-zero element. Since independent sources are necessarily uncorrelated, we would deny independence if the correlation coefficients of the estimated sources are significantly different from zero.

5.1. Estimated Sources Being Mutually Independent

We repeat the simulation 100 times to study the behavior of the test statistic on mutually independent sources. Computed test statistic as well as statistical inference drawn at $\alpha = 5\%$ for $n = 50, 200$ and 1000 are shown in Table 3. The second and third columns jointly gives the interval where 95% of the computed A_n lies. This can be contrasted with the critical values of A_n as shown in the next two columns [critical values can also be looked up in Table 1(B)]. If the computed A_n is not significant at the 95% confidence level, we reject H_1. This implies sources are independent. The $\hat{\alpha}$ in the last column represents the empirical Type I error which is the percentage H_0 erroneously rejected.

5.2. Estimated Sources Not Independent

We repeat the simulation 100 times to study the behavior of the test statistic on dependent estimated source signals. The 100 dependent samples are mostly synthetic, with about 20% due to unsuccessful attempts by some ICA algorithms. Computed test statistic and statistical inference drawn for $n = 50, 200$ and 1000 are shown in Table 4. The second and third columns gives the boundaries outside which 95% of the computed A_n lies. The $\hat{\beta}$ in the last column represents the empirical Type II error rate which is the percentage H_0 is erroneously accepted.

5.3. Tradeoff Between Type I and Type II Errors

By comparing the last column of Table 4 and that of Table 3, we can see the inevitable tradeoff between Type I error α and Type II error β. With α set at 5%, Type I error is smaller and empirically only accounts for about 2 out 100 simulations. However, this indirectly causes the Type II error to be greater than 5%. Since we cannot minimize Type I and Type II errors simultaneously, we would normally choose greater values for α if we regard the consequence of making wrong inference due to Type I error is less significant than that of Type II error. In other words, if we want to be more stringent on accepting independence.

6. CONCLUSION

Testing independence of estimated source signals is a vital task in real application because source signals are truly blind in the sense that both the source signals and the mixing matrix are not known. This paper illustrates how testing mutual independence between

sources can be effected via testing multivariate uniformity. In particular, several issues regarding adaptation of the uniformity test for independence are discussed.

7. REFERENCES

[1] P. Comon, "Independent component analysis - a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[2] N. Murata, "Properties of the empirical characteristic function and its application to testing for independence," *Proceedings of Third International Conference on Independent Component Analysis and Blind Signal Separation*, pp. 19–24, 2001.

[3] S. Shimizu and Y. Kano, "Examination of independence in independent component analysis," *Proceedings of 2001 International Meeting of the Psychometric Society*, 2001.

[4] R. B. D'Agostino and M. A. Stephens, *Goodness-of-Fit Techniques*, New York and Basel: Marcel Dekker, 1986.

[5] J. Liang, K. T. Fang, F. J. Hickernell, and R. Li, "Testing multivariate uniformity and its applications," *Mathematics of Computation*, vol. 70, no. 233, pp. 337–355, 2000.

[6] J. Neyman, ""smooth" test for goodness of fit," *Journal of American Statistical Association*, vol. 20, pp. 149–199, 1937.

[7] S. Pearson, E., "The probability transformation for testing goodness of fit and combining independenct tests of significance," *Biometrika*, vol. 30, pp. 134–148, 1939.

[8] F. J. Hickernell, "A generalized discrepancy and quadrature error bound," *Mathematics of Computation*, vol. 67, pp. 299–322, 1998.