Statistical Independence, Measures and Testing

***Y. Unnisa[1], D. Tran[1] and F. Huang[1]**

[1] College of Engineering and Science, Victoria University, FP Campus, PO Box 14428, MCMC 8001, Australia.

*Corresponding author: Danh.Tran@vu.edu.au

Abstract

Independent Component Analysis has recently been employed in structural damage detection and blind source separation to extract source signals and the unmixing matrix of the system from response signals. This novel method relies on the assumption that source signals are statistically independent. This paper looks at statistical independence, its measures and testing procedures. First the concepts of kurtosis, negentropy and mutual information are reviewed, followed by Bakirov's measures of coefficient of statistical independence and distance correlation between two signals coupled with Hypothesis testing to avoid Type I and Type II error. Bakirov's tests are nonparametric, simple to implement and do not require any approximation. Algorithms developed by Bakirov and associates to test the statistical independence of two arbitrary signals are reviewed. A case study using signals commonly found in vibration testing showed that Bakirov's tests are both reliable and rigorous. They are then applied to investigate the effects of corrupted signals by various forms on the statistical independence and performance of fastICA, a popular independent component analysis algorithm.

Keywords: Statistical independence, Bakirov's dCov test, Independent component analysis, Structural damage detection, Multivariate statistics, Package "energy", fastICA.

Introduction

Independent Component Analysis (ICA) is fundamentally a blind source separation method that seeks to separate underlying components from available data whether the data are in the form of sounds, images, vibration responses or financial share prices. Since 1990s, Independent Component Analysis has been of great interest to researchers in diversified areas of statistics, medical imaging, telecommunication and structural damage detection(Comon and Jutten, 2010, Hastie et al, 2008, Hyvärinen et al, 2001, Zang et al 2004). Essentially, ICA relies on response data collected by sensors, called *mixture signals*, and the assumption that the independent component sources, called *source signals,* are statistically independent, to extract the unknown source signals. Most of the studies require that there are as many sensors as there are independent components and that the system behaves linearly, but non-linear behavior and both under-determined and over-determined cases have also been solved. A well known case study is the so called cocktail party problem: identify speech by two speakers in a room by using sounds recorded by two microphones. A demonstration is given on http://research.ics.aalto.fi/ica/cocktail/cocktail_en.cgi.

ICA assumes that there is a relationship between **S**, the vector represents *source signals*, or underlying components and **X**, the vector represents *mixture or response signals* of the system to the source signals. In the simplest form, the relationship is linear and can be expressed as: **X** = **AS,** where **X** is available from sensors output, **A** is called the *mixing matrix.* **W,** the inverse of **A** is called the *unmixing matrix*. While both **A** or **W** and **S** have to be determined, ICA seeks the optimum solution out of all possible **W** such that the statistical independence of **S** is maximized. Naturally the product of **A** and **W** must be the identity matrix. In statistics, ICA is considered as supervised learning which includes principal component analysis and factor analysis. It is also connected to the technique of projection pursuit in multivariate statistics (Hastie et al, 2008). This has led to many novel methods of medical diagnosis of neurology and image processing. The restriction that has been stated by Hyvärinen et al (2001) is that not both variables are normal or Gaussian random signals. First, let us look at the concept of statistical independence (SI) and different measures and tests to evaluate statistical independence (SI).

Statistical Independence

Consider two scalar variables X and Y, X is said to be independent of Y if knowing the value of Y does not give any information on the value of X. This conceptual definition leads to the use of probability density function (pdf), a normalized histogram, of an event. When two events are studied, conditional probability, P(B|A), is defined as the probability that event B occurs given that event A occurs; and joint probability, P(A&B) is defined as the probability of both A and B occur. They are related by the rule P(B|A) = P(A&B)/P(A).

Two events are statistically independent if P(B|A) = P(B). It then follows that if A and B are independent: P(A&B) = P(A).P(B). The joint probability can be found by constructing a contingency table, however it should be noted that marginal probabilities can be found from joint probability but the reverse is not true except in the case of statistical independence. This leads to the notion that two scalar variables X and Y are statistically independent if and only if their jpdf is a product of their individual pdf which are also called marginal pdf:

$$p_{XY}(x,y) = p_X(x)\,.\,p_Y(y) \tag{1}$$

In Eq. 1, x and y are particular values of variable X, Y respectively.

Note that in Eq.1, cumulative distribution functions can replace the respective probability density functions, as so do expected values of absolutely integrable functions of variables, including positive powers of x and y:

$$E\{g(x).h(y)\} = E\{g(x)\}.\,E\{h(y)\} \tag{2}$$

$$E\{x^p\, y^q\} = E\{x^p\}\,.\,E\{y^q)\} \tag{3}$$

Where operator E stands for expected value, p and q are positive integers. It follows from Eq. 3 that SI is more stringent requirement than un-correlatedness, as un-correlatedness requires only E{x.y} = E{x}. E{y}, i.e only for the case that both p and q equal 1. Thus statistical independence implies un-correlatedness but the reverse is not true, except for normal or Gaussian random variable. A simple example is given by Stone (2004), in which two simple pendulums swinging 90^0 out of phase, x = cos(t), y = sin(t), giving correlation coefficient of zero, hence x and y are uncorrelated but they are statistically identical. At the same time, variables describing physically independent phenomena are intuitively thought to be statistically independent but it is not generally true.

Statistical Independence can also be defined in terms of characteristic functions of X and Y and their joint characteristic functions, where characteristic function of X is the inverse Fourier transform of its pdf and jpdf respectively, i.e. $f_X(t) = E\{e^{itX}\}$ and $f_{XY}(t,s) = E\{e^{i(Xt+Ys)}\}$. Note that characteristic functions are complex. In a similar fashion as using pdf and jpdf: X and Y are statistically independent if:

$$f_{XY}(t,s) = f_X(t).\, f_Y(s) \tag{4}$$

In most engineering applications variables are obtained from random process without further knowledge of the joint distribution, hence the jpdf cannot be determined from marginal pdfs, unless statistical independence is assumed or implied.

It must be noted that testing of the equality of the two sides of either of Eq. 1-4 highlights the basic concept of statistical hypothesis testing: a test must have hypotheses, the null and alternative hypothesis, a corresponding statistic and a measure of the reliability of the test. In other words the testing of Eq. 1 itself must be perceived in a probabilistic sense, not in a deterministic sense. This is to ensure not to commit Type I (rejecting true null hypothesis) and Type II error (accepting false null hypothesis).

Probably the first paper on statistical independence was due to Wilks (1935). Most researchers of ICA argue that the mixtures, as a consequence of Central Limit Theorem, would be more gaussian than the sources. As a consequence, a heuristic assumption is that the sources would be more non-Gaussian, hence the objective is seeking sources as variables of maximum non-Gaussianity, effectively using non-Gaussianity as a measure of statistical independence (Hyvärinen et al, 2001). Non-Gaussianity of a variable can be measured by kurtosis and negentropy. Kurtosis is defined as kurt(x) = $E\{x^4\} - 3(E\{x^2\})$ i.e. a normalized version of fourth moment of statistical distribution to make kurtosis of a normal or Gaussian random variable to be zero. Although simple to calculate, kurtosis is sensitive to outliers. The concept of entropy in Thermodynamics, representing the degree of being unstructured, unorganized, unpredictability, is also popular in Theory of Information. For a distribution Y, entropy of a variable is defined in terms of probability density function (pdf) as $H(y)= -\int p(y)\log p(y)dy$. Negentropy J is then defined as $J(y) = H(y_{Gauss}) - H(y)$, where y_{Gauss} is a Gaussian random variable of the same covariance matrix as y, which is shown by Information Theory to have the largest entropy among all random variables of equal variance. Thus negentropy is always non- negative. It is more involved to compute negentropy than kurtosis, and like kurtosis, it refers to only one variable and would fail as a measure of independence when one variable is a multiple of the other.

In most engineering applications, the variable has a finite number of values, as a consequence kurtosis of variables, even of the same distribution model, would depend heavily on how many elements are taken into account. As an example a variable was obtained by the Gaussian random generator in Matlab to yield a variable X of 1,000 elements, kurtosis was then found for varying number of elements from 100 to 1000. A typical result is shown in Table 1.

Table 1: Kurtosis of variables of varying number of elements from a normal (Gaussian) random variable

Element no	100	200	300	400	500	600	700	800	900	1000
kurtosis	-.3049	-.1041	-.1139	.0774	.0048	.2558	.2273	.1708	.2210	.1999

A more rigorous concept is mutual information of two variables X and Y defined as:

$$I(X,Y) = \sum_{y\in Y}\sum_{x\in X} p_{XY}(x,y)\log\left(\frac{p_{XY}(x,y)}{p(x)p(y)}\right) \quad (5)$$

It can be seen that mutual information is zero when the two variables are independent and ICA aims to minimize the mutual information among candidates of the source signals. Hyvärinen et al (2001) argued that this approach gives rigour to the more heuristic approach of using kurtosis and negentropy and is equivalent to method based on maximum likelihood estimation. However this measure of statistical independence also requires the knowledge of jpdf. Mutual information can be defined in terms of Shannon entropy, which can be further estimated (Comon and Jutten, 2010).

Bakirov's measures of statistical independence

Bakirov and his associates published two papers, Bakirov et al (2006), Szekely et al (2007) addressing the needs to have measure of statistical independence that are non-parametric, that is independent of the statistical model that one has to assume otherwise. Such a measure has to be practical to implement and conform to requirements of statistical hypothesis testing: null and alternative hypothesis, a test statistic and a confidence indicator of the test.

1. *Coefficient of independence*

Bakirov, Rizzo and Szekely proposed a statistic I_n based on the idea of independence coefficient I, defined in terms of characteristic functions:

$$I = \frac{\left\| f_{XY}(t,s) - f_x(t) f_Y(s) \ \right\|}{\left\| \sqrt{(1-|f_X(t)|^2)(1-|f_Y(s)|^2)} \ \right\|} \tag{6}$$

I_n itself is defined for a finite subset of variable of n elements, based on various Euclidean norms, or "distances" of distributions of X, Y and of their joint distribution Z, hence does not require the joint characteristic function as *I*. However, the authors proved that in the limit, I_n tends to *I* and $0 \le I_n \le 1$, where the sublimit 0 corresponds to statistical independence.

Further, it is shown that for all confidence level α below 0.215, the null hypothesis H_0 that X and Y are independent is rejected when $\sqrt{n} I_n \ge \phi^{-1}(1-\alpha/2)$ where ϕ^{-1}*is* the inverse function of the cumulative distribution function of the standard normal distribution. This assertion would yield a parameter indicating the strength of the hypothesis testing, normally given by the p-value of the hypothesis testing. It is normally accepted that p-value less than 0.05, H_0 would be rejected. The calculation of I_n is computing extensive for large n.

2. *Distance of covariance*

Szekely, Rizzo and Bakirov (2007) proposed the concept of distance covariance, *dCov* (X,Y) and distance correlation, *dCor* *R*(X,Y), defined respectively as:

$$\nu^2(\mathrm{X, Y}) = \|\mathrm{f}_{XY} - \mathrm{f}_X \mathrm{f}_Y\|^2 \tag{7}$$

$$R^2(X,Y) = \frac{\nu^2(X,Y)}{\sqrt{\nu^2(X)\nu^2(Y)}} \tag{8}$$

It can be seen from Eq. 7 that dCov is directly related to the definition of statistical independence. Further, the authors proved that the right hand side of Eq. 7 does not need information on the joint characteristic function and can be calculated as the limit of:

$$\nu_n^{\ 2}(\mathrm{X, Y}) = \mathrm{S}_1 + S_2 - 2S_3 \tag{9}$$

Where S_1, S_2 and S_3 can be calculated in terms of Euclidean norms related to distributions of X, Y. Similar hypothesis testing with statistic $n\nu_n^{\ 2}(\mathrm{X, Y})$ and p-value are also proposed.

3. *Implementation in R language*

The authors proposed two tests called *mvI.test* and *dcov.test*. Both tests use the null and alternative hypotheses H_0: p(x,y) = p(x).p(y), H_1: p(x,y) ≠ p(x).p(y). They are implemented as options in the module *indep.test* of the package *"energy'"* developed by Rizzo and Szekely in R language. The mvI.test corresponds to the coefficient of independence and takes longer than the dCov.test, as many times as the number of elements which can be in thousands or more. These tests yield the *p-value* of the null hypothesis test and it is widely accepted that H_0 should be rejected if p-value < 0.05. It should be noted that in statistical hypothesis test, p-value is viewed as a measure of the strength of the hypothesis test.

In this paper, Bakirov's dCov test is used to evaluate the statistical independence of source signals, measured by p-value of the test, before sending them to evaluate performance of ICA or to act as excitation signals in vibration testing or finite element simulation.

Statistical independence testing of common signals used in vibration

In this test, the interest is statistical independence of various excitation signals commonly used in vibration testing. Source signals of 1,000 elements were generated in Matlab, except that impact force signals were obtained in a vibration impact hammer test. These signals were then paired and tested for statistical independence by Bakirov's dCov test. Certainly if one signal is an exact copy or a multiple of the other, no matter what kind of signal, they would be tested dependent. The results are reported in Table 2. Typical plots of two signals for the case of sine-sawtooth pair and sinusoidal function of different frequencies (and also amplitude and phase) are shown in Figure 1 and 2.

Table 2: Statistical Independence of pairs of signals of 1,000 elements

Name of signal pair	dCov p-value
Unirandom, Unirandom (gemerated at different times)	0.660
Unirandom, Sine	0.415
Uirandom, Impact	0.635
Impact, Sine	0.5
Sine, Sine of different frequency	0.76
Sine, periodic Sawtooth of different frequency	0.965
Gaussian random, Gaussian random	0.815
Impact, Impact sampled at different points of structures	0.015
Unirandom (u), 5*u	0.005

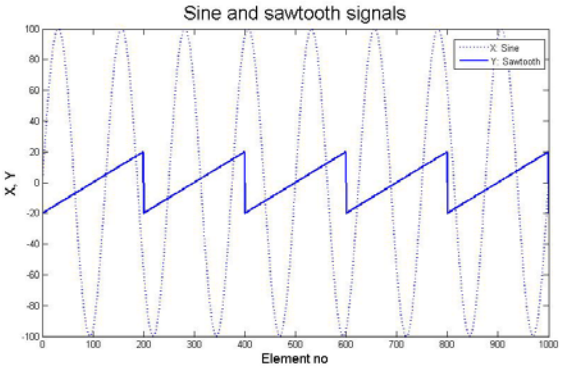

Figure 1: Sine and sawtooth signals

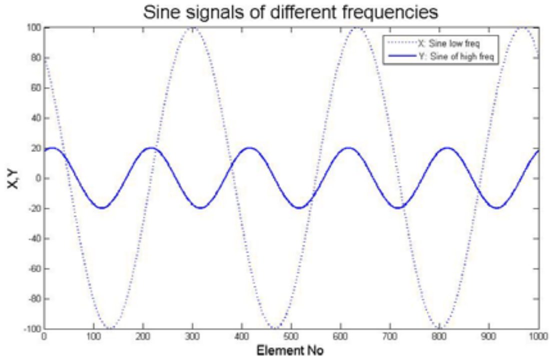

Figure 2: Sine signals of different frequencies

Inspection of Table 2 indicates that there is a variety of combination of different signals that would be statistical independent, except when one is a multiple of the other, or both are impact signals obtained from the same hammer tip-structure impact tests even if they were sampled at different points of the structure. It should be noted that as far as statistical independence is concerned, for sinusoidal signals, difference in frequencies is important whereas difference in amplitude or phase are not.

Statistical independence testing of corrupted signals

1. *One signal partially corrupted by the other signal*:

Two uniform random signals were generated in Matlab, called S1 and S2, of different ranges, each of 1000 elements. They are plotted against each other in Figure 3, showing the random nature of these two sources and their fast changing. The statistical independence of these signals was tested, giving p-value of 0.425, indicating that they are statistically independent. To avoid crowding, only the first 100 elements of S1 and S2 are plotted versus element number, as shown in Figure 4.

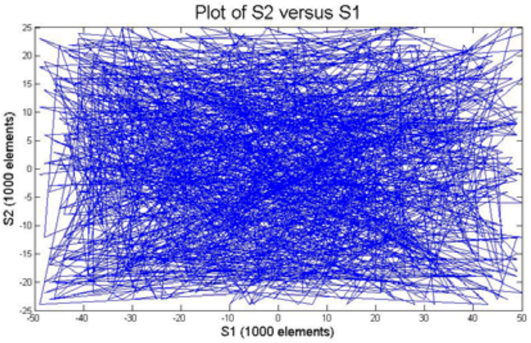

Figure 3: Uniform random source S2 versus S1

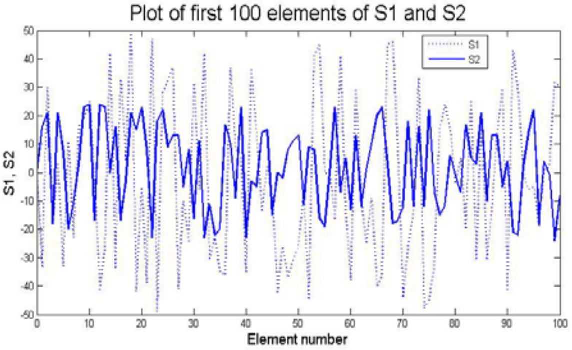

Figure 4: Plot of first 100 elements of S1 and S2

Next S1 was kept unchanged, S2 was changed by a varying percentage e% of the source signal S1. The new S2 is designated S2*, i.e. S2*= S2 + e%.S1. These new sets of signals S1 and S2* were then tested for statistical independence by Bakirov's dCov test. The following values of e% were investigated: 1, 2, 3, 4, and 10. The results are reported in Table 3,

Table 3: Effect of e% corruption of one signal on the other

e%	0	1	2	3	4	5	10
p-value of dCov test	0.425	0.555	0.43	0.185	0.045	0.005	0.005

It can be seen from Table 3 that in this case, Bakirov's dCov test of SI is very stringent: an addition of only 4% of S1 to S2 would make them not independent,

2. *Effect of random noise on statistical independence*

In this test, S1 was in the form of a sine wave and S2 was a sawtooth wave of equal amplitude of 1.00, as shown in Figure 5. They were then corrupted by Gaussian random noise of increasing amplitude of 0.05, 0.10, 0.15, 0.20. A plot of corrupted signals at amplitude of 0.10 is shown in Figure 6. The signals are tested for statistical independence in a similar fashion. The results are shown in Table 4.

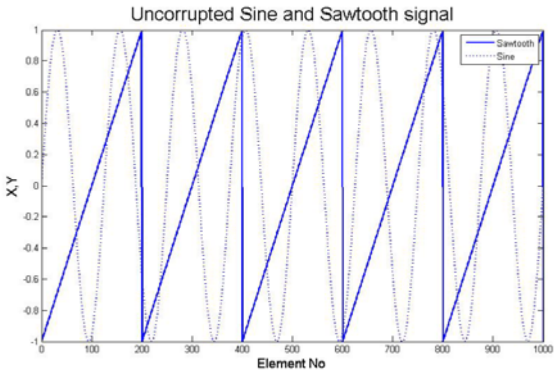

Figure 5: Sine and sawtooth signals

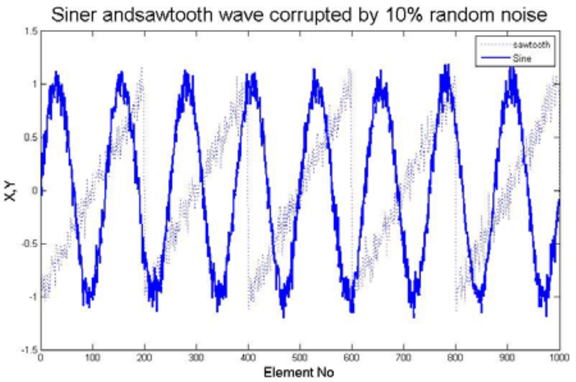

Figure 6; Corrupted sine and sawtooth signals by 10% gaussian random noise

Table 4: Effect of corruption of both signals by gaussian random noise

e% random noise	0	5	10	15	20
p-value dCov test	1.00	1.00	0.965	0.855	0.510

It was found that for the case of similar amplitude signals, Gaussian random noise, commonly exhibited by equipment used in vibration testing, does decrease the p-value of dCov test, but signals were still independent at 20% noise.

3. Evaluation the performance of Independent Component Analysis

As previously mentioned, Independent Component Analysis uses statistical independence as the objective function in searching for the blind sources from measured mixture signals. One popular algorithm is *fastICA*. Available in R, Matlab, C++ and Python programming, fastICA was developed by Marchini, Heaton and Ripley and can be downloaded from http://research.ics.aalto.fi/ica/fastica/. Basically it employs an approximation of negentropy as the objective function in searching of the unmixing matrix **W** under the constraints that **W** is an orthonormal matrix after the data has been centered, normalized and whitened. As the name implies it is a very fast algorithm, using fixed point iteration scheme for maximizing negentropy. It should be noted that the output of fastICA (source signals **S**, matrices **A** and **W**) are ambiguous as far as sign, scale and order are concerned. Here, the performance of fastICA was judged by the equality of **A*W** with the identity matrix of the same order. The signals used are S1 and S2* in Table 3. They were multiplied by a chosen **A** to yield the mixture signals which were then passed to fastICA for processing. The results are reported in Table 5, where **A0** is the mixing matrix corresponding to zero e%. It can be seen from Table 5 that up to adding 3% of S1 to S2, fastICA performed satisfactorily as far as the criterion of A*W = I is concerned, as expected. This equality is still satisfied at 4% but the mixing matrix obtained at this p% value is very different from the previous ones. This is further highlighted by inspecting the values of the ratio of the determinants of **A** at e% cases to that of the 0% case which was designated as **A0**. At 4% the ratio was 0.0048 instead of 1. At 10%, p-value was 0.005, fastICA failed to give the complete solution and no results reported.

Conclusions

The notion of statistical independence is very important in the area of blind source separation, including independent component analysis. It is shown that the non-parametric tests developed by Bakirov and associates, especially dCov test, provide a good measure of statistical independence. It was found that many signals commonly used as excitation sources in vibration testing are statistically independent, except when one is a multiple of the other, sinusoidal functions of the same frequency and impact signals sampled between the same hammer tip-structures in impact tests. The test was used to investigate effects of various sources of corruption on statistical independence: corruption of one signal by a small percentage of the other can affect enormously the statistical independence while corruption by random noise on both signals can be tolerated to a high level. It was also found that the statistical independence measured by p-value in dCov test is related to performance of fastICA, a popular package of ICA. It is recommended that Bakirov’ measures of statistical independence should be incorporated in an independent component analysis algorithm.

Table 5: Results of effects of statistical independence on performance of *fastICA*

e%	**dCov p-value**	**Mixing matrix A**	**Unmixing matrix W**	**A*W**	**detA/detA0**
0	0.425	-8.33133 -7.69348 -18.4173 7.9057	-0.03809 -0.03707 -0.08873 0.04014	1.0000 0.0000 0.0000 1.0001	1
1	0.555	7.865551 -8.01238 -7.50953 -18.7385	0.090281 -0.0386 -0.03618 -0.0379	1.0000 0.0001 -0.0001 1.0000	1
2	0.43	8.023941 7.693739 18.73223 -7.90597	0.03809 0.037068 0.090251 -0.03866	1.0000 0.0000 -0.0000 1.0000	1
3	0.185	-7.87029 -7.69384 -18.8896 7.906206	-0.03809 -0.03707 -0.09101 0.037919	1.0000 0.0000 -0.0000 1.0000	1
4	0.045	0.006301 0.99998 0.99998 -0.0063	0.006301 0.99998 0.99998 -0.0063	1.0000 0.0000 0.0000 1.0000	0.0048

References

Bakirov, N. K., Rizzo, M. L. and Szekely, G. J. (2006), A multivariate non-parametric test of independence, *Journal of Multivariate Analysis*, 97 pp.1742-1756.

Comon, P. and C. Jutten, C. (2010), Handbook of Blind Source Separation, Independent Component analysis and Applications, Academic Press, Amsterdam.

Hastie, T., Tibshirani, R. and Friedman, J. (2008), The elements of statistical learning, data mining, inference and prediction, Second Edition, Springer.

Hyvärinen , A., Karhunen, J. and Oja, E. (2001), Independent Component Analysis, Wiley, New York.

Stone, J. V. (2004), Independent Component Analysis, a tutorial introduction, The MIT Press, Massachusetts.

Szekely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007), Measuring and testing dependence by correlation of distances, *Journal of The Annals of Statistics,* 35, 6 pp. 2769–2794.

Zang, C., Friswell, M. I., and Imregun, M. (2004), Structural Damage Detection using Independent Component Analysis, *Structural Health Monitoring*, 3(1) pp. 69-83.

Wilks, S. S. (1935), On the independence of k sets of normally distributed statistical variable, *Econometrica,* Vol. 3, No. 3 pp. 309-326.