NEW CORRELATION COEFFICIENT FOR DATA ANALYSIS

Dragos FALIE[1], Livia DAVID[2]

[1]Politechnical University of Bucharest, Romania, 1-3 Iuliu Maniu, 030791, Bucharest, Romania, E-mail : Dragos.Falie@IEEE.org,
[2]University of Agricultural Sciences and Veterinary Medicine, Bucharest, 59 Marasti, sector 1, 011464, Bucharest, Romania, E-mail:Livia_David@hotmail.com, Phone: +40722602064

Corresponding author email: dragos.falie@ieee.org

Abstract

The proposed correlation coefficient better characterize the statistical independence of two random variables that are a linear mixture of two independent sources. This correlation coefficient can be calculated with analytical relations or with the known algorithms of independent components analysis (ICA). The value of the correlation coefficient is zero when the random variables are a statistically independent and it is one when these are fully dependent.

Keywords: *blind separation of sources, correlation, ICA, statistical independence*

INTRODUCTION

The dependences between two random variables and is represented generally by a correlation relation and the commonly used is the Pearson correlation coefficient[1]:

$$\rho\left(\tilde{\mathbf{x}}_1,\tilde{\mathbf{x}}_2\right)=\frac{E\left[\left(\tilde{\mathbf{x}}_1-E\left[\tilde{\mathbf{x}}_1\right]\right)\cdot\left(\tilde{\mathbf{x}}_2-E\left[\tilde{\mathbf{x}}_2\right]\right)\right]}{\sigma(\tilde{\mathbf{x}}_1)\cdot\sigma(\tilde{\mathbf{x}}_2)} \qquad (1)$$

where E is the expectation operator and σ is the standard deviation of a random variable :

$$\sigma(\tilde{\mathbf{x}})=\sqrt{E\left[\left(\tilde{\mathbf{x}}-E\left[\tilde{\mathbf{x}}\right]\right)^2\right]} \qquad (2)$$

The correlation coefficient (1) has a simpler relation:

$$\rho\left(\tilde{\mathbf{x}}_1,\tilde{\mathbf{x}}_2\right)=E\left[\mathbf{x}_1\cdot\mathbf{x}_2\right] \qquad (3)$$

if the random variables and are normalized:

$$\mathbf{x}_1=\frac{\tilde{\mathbf{x}}_1-E\left[\tilde{\mathbf{x}}_1\right]}{\sigma(\tilde{\mathbf{x}}_1)},\quad \mathbf{x}_2=\frac{\tilde{\mathbf{x}}_2-E\left[\tilde{\mathbf{x}}_2\right]}{\sigma(\tilde{\mathbf{x}}_2)} \qquad (4)$$

The simplest situation is when $\mathbf{x}_1$, $\mathbf{x}_2$ are a linear mixture of two statistically independent normalized random variables $\mathbf{s}_1$, $\mathbf{s}_2$ named sources:

$$\mathbf{x}_1=a_{11}\cdot\mathbf{s}_1+a_{12}\cdot\mathbf{s}_2,\quad \mathbf{x}_2=a_{21}\cdot\mathbf{s}_1+a_{22}\cdot\mathbf{s}_2 \qquad (5)$$

In this case the correlation coefficient between $\mathbf{x}_1$, $\mathbf{x}_2$ is:

$$\rho\left(\mathbf{x}_1,\mathbf{x}_2\right)=a_{11}\cdot a_{21}+a_{12}\cdot a_{22} \qquad (6)$$

Assuming that the unit vectors along the *x,y* axis corresponds to $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{x}_1$, $\mathbf{x}_2$ are given by (5) then $\mathbf{x}_1$ can be represented in a □ ^{2} space by the vector $[a_{11}, a_{12}]$ and $\mathbf{x}_2$ by the vector $[a_{21}, a_{22}]$. With this representation the correlation coefficient (6) can be represented geometrically as the scalar product between $\mathbf{x}_1$ and $\mathbf{x}_2$.

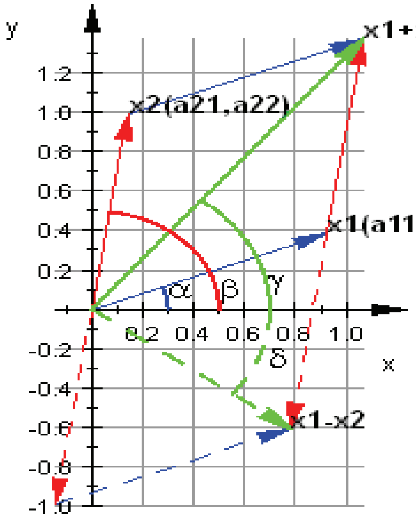

Fig. 1. The dependence of $\mathbf{x}_1$, $\mathbf{x}_2$ on $\mathbf{s}_1$, $\mathbf{s}_2$. On the *x* ax that corresponds to $\mathbf{s}_1$, the coefficients a_{11}, a_{12} are represented. On the *y* ax that corresponds to s $\mathbf{s}_2$, the coefficients a_{21}, a_{22} are represented.

Due to the fact that $\mathbf{x}_1$, $\mathbf{x}_2$ are normalized then $a_{11}^2+a_{12}^2=1$ and $a_{21}^2+a_{22}^2=1$. In this case the relations (5) can be rewritten as:

$$\begin{aligned}\mathbf{x}_1&=\mathbf{s}_1\cos(\alpha)+\mathbf{s}_2\sin(\alpha)\\ \mathbf{x}_2&=\mathbf{s}_1\cos(\beta)+\mathbf{s}_2\sin(\beta)\end{aligned} \qquad (7)$$

where the α, β are the angles formed by $\mathbf{x}_1$, $\mathbf{x}_2$ with the x ax respectively.
Using (7) the correlation coefficient takes a very simple trigonometric form:

$$\rho(\mathbf{x}_1, \mathbf{x}_2) = \cos(\alpha - \beta) \qquad (8)$$

In the case when both $\mathbf{x}_1$, $\mathbf{x}_2$ have a Gaussian distribution or any one of the coefficients a_{11}, a_{12} a_{21} and a_{22} equals to zero, then the absolute value of the correlation coefficient measure the statistical dependence between the random variables $\mathbf{x}_1$, $\mathbf{x}_2$. In this case if the correlation coefficient is zero then $\mathbf{x}_1$, $\mathbf{x}_2$ are statistically independent. In the other cases the correlation coefficient may not correctly show the statistical dependence between $\mathbf{x}_1$ and $\mathbf{x}_2$.
For example the Pearson's correlation coefficient expressed by Eq. (8) is zero in the case when the random variables are "orthogonal":

$$\alpha - \beta = \frac{\pi}{2} + k \cdot \pi, \quad k \in \mathbb{Z} \qquad (9)$$

In this case the variables $\mathbf{x}_1$, $\mathbf{x}_2$ are not statistical independent if in (8) $\alpha \neq +k\pi/2$ and quite dependent in the particular case when $\alpha = \pi/4$ and $\beta = \alpha + \pi/2$:

$$\mathbf{x}_1 = \frac{\sqrt{2}}{2}(\mathbf{s}_1 + \mathbf{s}_2), \quad \mathbf{x}_2 = \frac{\sqrt{2}}{2}(\mathbf{s}_1 - \mathbf{s}_2) \qquad (10).$$

MATERIAL AND METHOD

The random variables $\mathbf{x}_1$, $\mathbf{x}_2$ given by (5), are independent, when $a_{11}a_{12} = 0$ and $a_{21}a_{22} = 0$. In this case, but not only, the Pearson correlation coefficient (6) is zero. It would be therefore useful to provide an indicator, which is different from zero when the variables $\mathbf{x}_1$ and $\mathbf{x}_2$ are dependent but the Pearson coefficient is zero.
The new correlation coefficient that we propose is defined with the following formula:

$$R(\mathbf{x}_1, \mathbf{x}_2) = |a_{11} \cdot a_{21}| + |a_{12} \cdot a_{22}| \qquad (11)$$

The value of R is zero only when the random variables (5) are statistical independent and one when these are fully dependent. It has be noted with the Latin letter R similar with the correlation coefficient that is usually noted with the Greek letter ρ.
By using (7) the correlation coefficient R can be expressed as:

$$R(\mathbf{x}_1, \mathbf{x}_2) = \max\left\{|\cos(\alpha - \beta)|, \ |\cos(\alpha + \beta)|\right\} \qquad (12)$$

It can be noticed that the Pearson correlation coefficient expressed as in Eq.(8) is the same with R when:

$$|\cos(\alpha - \beta)| > |\cos(\alpha + \beta)| \qquad (13)$$

On Fig. 2 is represented the particular case when $\mathbf{x}_1$, $\mathbf{x}_2$ are orthogonal $\beta=\alpha+(2k+1)\pi/2$. In this case the Pearson's correlation coefficient (8) is zero but, R may vary from 0 to 1:

$$R(\mathbf{x}_1, \mathbf{x}_2) = |\sin(2 \cdot \alpha)|, \quad \beta = \alpha \pm (2k+1)\frac{\pi}{2}, \quad k \in \mathbb{Z} \qquad (14)$$

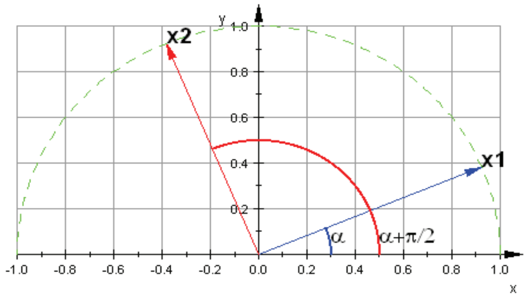

Fig. 2. The dependence of two orthogonal random variables $\mathbf{x}_1$, $\mathbf{x}_2$ on s1, s2. The x, y axis corresponds to s1, s2.

When $\alpha = \pi/4$ then R=1 and $\mathbf{x}_1$, $\mathbf{x}_2$ are in the most dependent situation. Other particular cases are when R=1/2 when $\alpha = \pi/12$ and R=√3/2 when $\alpha = \pi/6$.
The random variables $\mathbf{x}_1$, $\mathbf{x}_2$ are fully dependent and R=1 when:

$$\begin{aligned} &R(\mathbf{x}_1, \mathbf{x}_2) = \sin(\alpha)^2 + \cos(\alpha)^2 = 1, \\ &\beta = \pm\alpha + k \cdot \pi, k \in \mathbb{Z} \end{aligned} \qquad (15)$$

The correlation coefficient R can be calculated by Eq. (11) if $\mathbf{x}_1$, $\mathbf{x}_2$ are separated into independent components by using an independent components analysis (ICA) algorithm [2-4].
R may be calculated also with Eq. (12) in which case, what is needed is, to evaluate cos(α+β), cos(α-β) being known via(8).
To compute R with (12) is necessary to know the value of:

$$r(\mathbf{x}_1, \mathbf{x}_2) = \cos(\alpha + \beta) \qquad (16)$$

Analytical solution for $r(\mathbf{x}_1, \mathbf{x}_2)$ is:

$$\cos(t)^2 = \frac{k_{40} - 2k_{22} + k_{04}}{k_{40} + 2k_{22} + k_{04}} \qquad (17)$$

where:

$$\begin{aligned} &k_{40} = E\left[\mathbf{o}_1^4\right] - 3, \ k_{04} = E\left[\mathbf{o}_2^4\right] - 3, \\ &k_{22} = E\left[\mathbf{o}_1^{\,2}\mathbf{o}_2^{\,2}\right] - 1 \end{aligned} \qquad (18)$$

Eqs. (17), can be obtained only when the following condition is fulfilled:

$$E\left[\left(\mathbf{o}_1^2+\mathbf{o}_2^2\right)^2\right]-8=E\left[\mathbf{s}_1^4\right]+E\left[\mathbf{s}_2^4\right]-6\neq 0 \quad (19)$$

For a Gaussian source E[s4]= 3 and in this case if both sources $\mathbf{s}_1$, $\mathbf{s}_2$ are Gaussian (19) is not fulfilled. If for one of the sources E[$\mathbf{s}_1^4$]<3 and for the other E[$\mathbf{s}_2^4$]>3 such as (19) is not fulfilled then the solution cannot be calculated with (17). Also in the case when one of the sources are a mixture of two random variables such as E[$\mathbf{s}_1^4$]=3 and the other source is or not Gaussian but for it also E[$\mathbf{s}_2^4$]=3 then the solution cannot be computed with (17).

When (19) is fulfilled *R* can be calculated knowing *tan*(*t*) obtained with the Comon's relation [5] or with the alternative Comon's formula (ACF)[8,3,14]:

$$\tan(t)=\frac{2k_{22}}{k_{31}-k_{13}} \quad (20)$$

where:

$$k_{31}=E\left[\mathbf{o}_1{}^3\mathbf{o}_2\right],\quad k_{13}=E\left[\mathbf{o}_1\mathbf{o}_2{}^3\right] \quad (21)$$

The best results are obtained with the following relation:

$$\tan(2t)=\frac{4\left(k_{31}-k_{13}\right)}{k_{40}-6k_{22}+k_{04}} \quad (22)$$

The above relation is known as the approximate maximum likelihood (AML) estimator [10-12,5]. This relation can also be obtained by combining E[$\mathbf{o}_1\mathbf{o}_2^3$] and E[$\mathbf{o}_1^3\mathbf{o}_2$. Additionally the condition (19) need to be fulfilled.

RESULTS AND DISCUSSIONS

R corrects the Pearson's correlation coefficient only when all the coefficients a_{11}, a_{12}, a_{21} and a_{22} in (5) are different from zero. If one of these coefficients equal zero then the system (5) reduces to:

$$\mathbf{x}_1=\mathbf{s}_1,\quad \mathbf{x}_2=a_{21}\cdot\mathbf{s}_1+a_{22}\cdot\mathbf{s}_2 \quad (23)$$

and the two correlation coefficients gives the same result.

The correlation matrix *ρ* and *R* between the changing rates of different currency are presented in the Tab. 1 and 2 respectively. The difference between *ρ* and *R* is presented in Table 3.

A general remark is that there are enough cases where *ρ* has been corrected by *R* to justify the use of the new correlation coefficient. In this example the corrected correlation *R* has a greater value than *ρ*.

TABLE 1.CORRELATION MATRIX

	gold	**$ USA**	**€**	**£ UK**	**f Sw**	**$ Ca**	**$ Au**
gold	0.996	0.953	0.230	0.854	0.860	0.940	0.916
$ USA	0.953	0.996	0.196	0.833	0.820	0.964	0.877
€	0.230	0.196	0.996	0.562	0.479	0.085	0.187
£ UK	0.854	0.833	0.562	0.996	0.915	0.765	0.801
f Sw	0.860	0.820	0.479	0.915	0.996	0.785	0.878
$ Ca	0.940	0.964	0.085	0.765	0.785	0.996	0.927
$ Au	0.916	0.877	0.187	0.801	0.878	0.927	0.996

TABLE 2.CORRECTED CORRELATION MATRIX

	gold	**$ USA**	**€**	**£ UK**	**f Sw**	**$ Ca**	**$ Au**
gold	0.996	0.983	0.286	0.999	0.906	0.940	0.954
$ USA	0.983	0.996	0.279	1.000	0.996	0.964	0.977
€	0.286	0.279	0.996	0.562	0.479	0.085	0.249
£ UK	0.999	1.000	0.562	0.996	0.993	0.984	1.000
f Sw	0.906	0.996	0.479	0.993	0.996	0.999	0.969
$ Ca	0.940	0.964	0.085	0.984	0.999	0.996	1.000
$ Au	0.954	0.977	0.249	1.000	0.969	1.000	0.996

TABLE 3.THE DIFFERENCE BETWEEN THE TWO CORRELATION COEFFICIENTS

	gold	**$ USA**	**€**	**£ UK**	**f Sw**	**$ Ca**	**$ Au**
gold	0.000	-0.030	-0.056	-0.145	-0.046	0.000	-0.039
$ USA	-0.030	0.000	-0.083	-0.167	-0.176	0.000	-0.100
€	-0.056	-0.083	0.000	0.000	0.000	0.000	-0.062
£ UK	-0.145	-0.167	0.000	0.000	-0.078	-0.219	-0.199
f Sw	-0.046	-0.176	0.000	-0.078	0.000	-0.214	-0.090
$ Ca	0.000	0.000	0.000	-0.219	-0.214	0.000	-0.073
$ Au	-0.039	-0.100	-0.062	-0.199	-0.090	-0.073	0.000

As was expected there are also a lot of cases where the data structure has the simple form as in (23) which case the two correlation coefficients gives the same or very close results. For example in the 5th row of table 3 the dependence of the Canadian $ on gold, USA $ and € has a simple structure but, the dependence on the £ UK and Swiss franc is complex it impose the use of the new correlation coefficient.

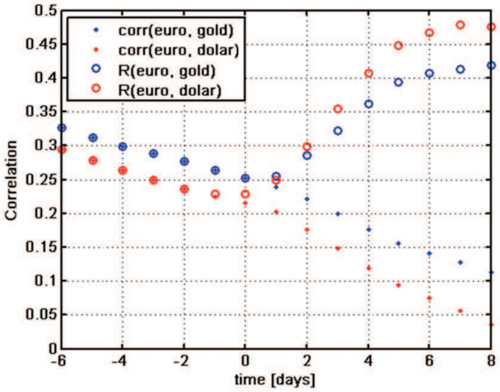

Fig. 3. The time dependence of the correlation coefficients ρ and R between € and the gold price and $ USA.

The time dependence of the correlation coefficients is represented in Fig. 3. One can observe that if the data of the gold price and USA $ are shifted in the past with 0…6 days the values of the two correlation coefficients are the same. This indicates that the data structure in these cases has the simple form(23).

The data structure changes if the same data (gold price and USA $) is shifted in the future with 1…8 days. This example shows that if the data structure is not priory known is better to use the new correlation coefficient.

CONCLUSIONS

The proposed correlation coefficient R corrects the Pearson relation and show the statistical dependence between two random variables that are a linear mixture of two independent sources. *R* can be calculated with analytical relations or can be obtained by ICA algorithms also. Beside the known relation to calculate *R* a new analytical relation has been proposed (17).

There are situations when the random variables does not satisfy the condition (19) and *R* needs to be calculated using ICA algorithms [7].

Even if to compute *R* is a little bit more complicate than to calculate the correlation coefficient the advantage to know it (*R*) are considerable. In many cases the random variables does not have the simple structure of (23) and the Pearson's correlation coefficient may give inaccurate values.

In economics, stock market [9] and other fields the dependences between different random variables cannot always be correctly evaluated with the correlation coefficient but *R* can easily clarify the problem.

REFERENCES

[1]Back, A.D. and A. S. Weigend, *A first application of independent component analysis to extracting structure from stock returns,* International Journal of Neural Systems, vol. 8, pp. 473-484, Aug 1997.

[2]Comon, P., *Separation of Stochastic Processes*, in Proceedings Workshop on Higher-Order Spectral Analysis, 1989, pp. pp. 174–179.

[3]Comon, P. *Separation of sources using higher-order cumulants*. Vol. 1152, 1989.

[4]De la Rosa, J.J.G. et al., *Higher order statistics and independent component analysis for spectral characterization of acoustic emission signals in steel pipes*, Ieee Transactions on Instrumentation and Measurement, vol. 56, pp. 2312-2321, Dec 2007.

[5]Harroy, F. and J.-L. Lacoume, *Maximum likelihood estimators and Cramer-Rao bounds in source separation*, Signal Processing, vol. 55, pp. 167-177, 1996.

[6]Hyvarinen, A., *Fast and robust fixed-point algorithms for independent component analysis*, Ieee Transactions on Neural Networks, vol. 10, pp. 626-634, May 1999.

[7]Hyvarinen,A., *The fixed-point algorithm and maximum likelihood estimation for independent component analysis*, *Neural Processing Letters,* vol. 10, pp. 1-5, Aug 1999.

[8]Lacoume, J.L. et al., *Statistiques d'Ordre Supérieur pour le Traitement du Signal*. Paris: Masson, 1997.

[9]Stuart, A. and J. K. Ord, *Kendall's Advanced Theory of Statistics*, sixth ed. vol. I. London: Edward Arnold, 1994.

[10]Zarzoso,V. et al., *Optimal pairwise fourth-order independent component analysis,* Ieee Transactions on Signal Processing, vol. 54, pp. 3049-3063, Aug 2006.

[11]Zarzoso, V. and A. K. Nandi, *Closed-form estimators for blind separation of sources - part II: Complex mixtures*, Wireless Personal Communications, vol. 21, pp. 29-48, Apr 2002.

[12]Zarzoso, V. and A. K. Nandi, *Closed-form estimators for blind separation of sources - part I: Real mixtures*, Wireless Personal Communications, vol. 21, pp. 5-28, Apr 2002.

[13]Zarzoso, V. et al., *A contrast for independent component analysis with priors on the source kurtosis signs,* Ieee Signal Processing Letters, vol. 15, pp. 501-504, 2008 2008.

[14]Zhang, Z. et al., *Blind source separation by combining independant component analysis with complex discrete wavelet transform,* 2007 International Conference on Wavelet Analysis and Pattern Recognition, Vols 1-4, Proceedings, pp. 549-554, 1924, 2007.