Statistical distance for random variables

D. FALIE
Faculty of Electronics, Telecommunications and Information Technology, University Politehnica of Bucharest, Bd. Iuliu Maniu 1-3, Bucureşti 030791, Romania

Two statistical distances for random variables, based on the standard deviation operator, and on the mean absolute deviation, are proposed. The numerical values obtained with these relations are the same if both random variables are Gaussian, and may differ in the other cases. A generalized mean deviation covariance and correlation coefficients are also defined. These new statistical operators can be used to detect the non-Gaussian random or a mixture of random variables. The generalized statistical operators are differently affected by outliers and from these on optimum operator can be choose. These new statistical operators have been used for gaze guidance evaluations and for scan path measurements.

(Received March 06, 2010; accepted July 14, 2010)

*Keywords:S*statistical distance, Non-gaussianity, Outlier, Generalized correlation, Generalized covariance

1. Introduction

The correlation coefficient generally used in statistics is the Pearson's product moment(1), where E is the expectation operator and σ is the standard deviation. The Eq. (1) can be rewritten as (2) where by X_n and Y_n are noted the normalized values of the random variable X and Y (3).

$$corr(X,Y) = \frac{E\left[\left(X - E[X]\right)\cdot\left(Y - E[Y]\right)\right]}{\sigma(X)\cdot\sigma(Y)} \tag{1}$$

$$corr(X,Y) = E\left[\frac{X - E[X]}{\sigma(X)}\cdot\frac{Y - E[Y]}{\sigma(Y)}\right] = E\left[X_n \cdot Y_n\right] \tag{2}$$

$$X_n = \frac{X - E[X]}{\sigma(X)}, \quad Y_n = \frac{Y - E[Y]}{\sigma(Y)} \tag{3}$$

The pseudo-metric function can be defined using the correlation between X and Y with Eq.(4). The statistical distance between the random variables is zero when these are fully correlated; when these are independent or uncorrelated the value of the distance reaches the maximum value of one.

$$d(X,Y) = 1 - \left|corr(X,Y)\right| \tag{4}$$

A function of the random variables X and Y must satisfy the following conditions to be considered a pseudo-metric function:

$$d(X,Y) \geq 0 \tag{5}$$

$$d(X,Y) = 0, \quad if \quad X = Y \tag{6}$$

$$d(X,Y) = d(Y,X) \tag{7}$$

$$d(X,Z) \leq d(X,Y) + d(Y,Z) \tag{8}$$

The function described by Eq. (4) satisfies the first three conditions but not the last condition called "subadditivity" or the triangle inequality. More precisely, Eq. (4) always does not satisfy the triangle inequality, and seems necessary to find a different correlation coefficient. Bradley has found a different correlation coefficient (6) [1] using the mean (average) absolute deviation (MAD) (5) (which is the average absolute deviation from the mean of a random variable X). The random variables u and v in Eq. (6) must satisfy the conditions (7). The values given by the Bradley correlation coefficient differ from those computed with the Pearson's relation (1) but this is less sensitive to outliers than the Pearson's correlation coefficient [2-4]. Unfortunately with the Bradley' correlation coefficient can not be defined a statistical distance using (4).

$$D(X) = E\left[\left|X - E[X]\right|\right] \tag{5}$$

$$\rho(u,v) = \frac{E(|u+v| - |u-v|)}{E(|u| + |v|)} \tag{6}$$

$$E(|u|) = E(|v|), \quad E(u) = E(v) = 0 \tag{7}$$

2. Generalized mean deviation

The relation of Pearson's correlation coefficient can be rewritten using the standard deviation operator as (8).

$$corr(X,Y) = \frac{1}{4}\left[\sigma^2\left(\frac{X}{\sigma(X)} + \frac{Y}{\sigma(Y)}\right) - \sigma^2\left(\frac{X}{\sigma(X)} - \frac{Y}{\sigma(Y)}\right)\right] \tag{8}$$

A different relation, but with similar proprieties, is obtained if instead of the standard deviation operator the

MAD is used. Other different expressions of the correlation coefficient can be obtained if the standard deviation is replaced with other operators.

The common property of the standard deviation and of the MAD is (9) or (10). We define the general expression of the generalized mean deviation (GMD) based on the function f_σ with the equation (11) if the condition (12) is fulfilled.

$$\sigma(a \cdot X + b) = |a| \cdot \sigma(X) \tag{9}$$

$$E\Big[\big|a \cdot X + b - E[a \cdot X + b]\big|\Big] = |a| \cdot E[X] \tag{10}$$

$$\sigma_f(X) = f_\sigma^{-1}\Big[E\Big[f_\sigma\big(|X - E[X]|\big)\Big]\Big] \tag{11}$$

$$\sigma_f(a \cdot X + b) = |a| \cdot \sigma_f(X) \tag{12}$$

As an example we compute the GMD of the power function (13). In the case when X is normal distributed random variable it can be represented as Eq. (14) where x is a normal random variable with zero mean and with a standard deviation equal to one. The GMD is given by Eq. or (16) where c_p is a constant that does not depend on the MAD or the standard deviation of the variable X it depends only on the distribution function of the random variable.

$$f_\sigma(|x|) = |x|^p, \quad f_\sigma^{-1}(|x|) = |x|^{\frac{1}{p}} \tag{13}$$

$$X = \sigma \cdot x + \mu, \quad \sigma = \sqrt{E\Big[\big(X - E[X]\big)^2\Big]}, \quad \mu = E[X] \tag{14}$$

$$\sigma_p(X) = \Big(E\Big[|X - E[X]|^p\Big]\Big)^{\frac{1}{p}} = \sigma \cdot \Big(\int_{-\infty}^{+\infty} f(x;0,1) \cdot |x|^p \, dx\Big)^{\frac{1}{p}} \tag{19}$$

$$f(x;0,1) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} e^{-\frac{x^2}{2 \cdot \sigma^2}} \tag{15}$$

$$\sigma_p(X) = \sigma \cdot c_p^{\frac{1}{p}} \quad c_p = \int_{-\infty}^{+\infty} f(x;0,1) \cdot |x|^p \, dx \tag{16}$$

If the distributed random variable X has an uniform distribution in the interval [-a, a] then the GMD is given by Eq. (17) .

$$\sigma_p(X) = \frac{a}{(p+1)^{\frac{1}{p}}} = \sigma(X) \cdot 3 \cdot (p+1)^{-\frac{1}{p}} \tag{17}$$

The logarithm function satisfy the condition (12) and a GMD can be defined using this function (19). The value of the constant $\sigma_{\log}$, when X has a normal distribution is given by (20) .

$$\begin{aligned} &\exp\Big(E\big[\log(|a \cdot X|)\big]\Big) = \exp\Big(E\big[\log(|a|) + \log(|X|)\big]\Big) = \\ &= \exp\Big(\log(|a|) + E\big[\log(|X|)\big]\Big) = |a| \cdot \exp\Big(E\big[\log(|X|)\big]\Big) \end{aligned} \tag{18}$$

$$\sigma_{\log}(X) = \sigma \cdot \exp(c_{\log}) \tag{19}$$

$$c_{\log} = \int_{-\infty}^{+\infty} f(x;0,1) \cdot \log(|x|) \, dx \tag{20}$$

For a normal distribution the GMD of the central absolute moments is Eq. (21), and the values of C_p are given by (22) and (23). The values of C_p for a uniform distributed random variable are given by (24), and this value decrease when p increase while, in the case of a normal distribution, these values increases.

$$\sigma_p(X) = \Big(E\Big[|X - E[X]|^p\Big]\Big)^{\frac{1}{p}} = C_p \cdot \sigma(X) \tag{21}$$

$$C_p = \left(\frac{(2 \cdot n)!}{2^n \cdot n!}\right)^{\frac{1}{p}}, \quad p = 2 \cdot n \tag{22}$$

$$C_p = \left(\sqrt{\frac{2}{\pi}} \cdot 2^n \cdot n!\right)^{\frac{1}{p}}, \quad p = 2 \cdot n - 1 \tag{23}$$

$$C_p = \left(\frac{1}{p+1}\right)^{\frac{1}{p}} \tag{24}$$

From Eq. (11) and (12) results that always we have (25), (26), because $\sigma(X)$ is a constant. In the particular case when f_σ is the power function (13) then $C_p(p)$ is a function on p and it is a functional transform of the random variable's distribution. $C_p(p)$ is the GMD of the normalized variable, and this function can be computed or determined. The $C_p(p)$ function is used to find a suitable value of p that will be used to detect changes of the distribution function, non-Gaussianity or to compute a corrected standard deviation. For example $C_p(1/4)$ can be used to detect non-Gaussianity instead of kurtosis.

$$\sigma_f(X) = \sigma(X) \cdot f_\sigma^{-1}\left[E\left[f_\sigma\left(\frac{|X - E[X]|}{\sigma(X)}\right)\right]\right] = \sigma(X) \cdot C_f \tag{25}$$

$$C_f = f_\sigma^{-1}\big[E[f_\sigma(x)]\big], \quad x = \frac{|X - E[X]|}{\sigma(X)} \tag{26}$$

For example we need to find a proper value of p to detect that the pseudo random variable `xp = random('poiss', 9, 10^6, 1)` generated with MatLab is non-Gaussian, or if its distribution have been changed by adding 1% of a normal or uniform distributed random variable with the same standard deviation.

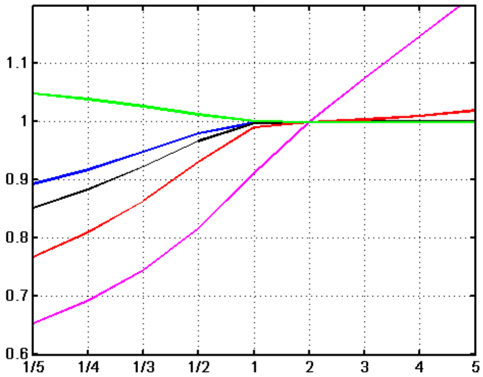

Fig. 1. The ratio of C_p(p,xp+xu/100)/C_p(p,xp+xn/100) is represented with green, C_p(p,xp)/C_p(p,xp+xn/100) with blue, C_p(p,xp)/C_p(p,xp+xu/100) with black, C_p(p,xp)/C_p(p,xn) with red, and C_p(p,xp)/ C_p(p,xu) with magenta. On the abscise are represented the values of p.

The values of $C_p(p)$ of the random variable xp, xn, xu, xp+xu/100, and xp+xn/100 are noted with $C_p(p,xp)$, $C_p(p,xn)$, $C_p(p,xu)$, $C_p(p,xp+xu/100)$, $C_p(p,xp+xn/100)$, where `xn = random('norm', 0, 3, 10^6, 1);xu = random('unif', -5.2, 5.2, 10^6, 1).` Annualizing the graphs in Fig.1 appears rather difficult to establish that the random variable xp is Gaussian or not using the forth-centered moment. This problem can be solved using the ratio $C_p(1/4,xp)/C_p(1/4,xn)$ of the GMDs, because this value differ with about 20% when the random variable has a Poisson distribution.

3. Generalized correlation coefficient

The generalized correlation coefficient is a generalization of Eq. (8) by using the GMD instead of the standard deviation. In the classical relation the two main expressions are raised to the power of two, but in the generalized expression these can be raised at any real power. Using the pseudo-normalized variables (28) the Eq. can be rewritten as Eq. (27). In a similar way can be defined a generalized expression of the covariance (29). The covariance equals with the generalized covariance for *p=2* and *q=2* , and for *q=2* and different p values the generalized covariance equals the covariance multiplied by a constant (30).

$$corr_{p,q}(X,Y)= \\ =\frac{1}{2^q}\left[\sigma_p^q\left(\frac{X}{\sigma_p(X)}+\frac{Y}{\sigma_p(Y)}\right)-\sigma_p^q\left(\frac{X}{\sigma_p(X)}-\frac{Y}{\sigma_p(Y)}\right)\right] \quad (32)$$

$$corr_{p,q}(X,Y)=\frac{1}{4}\left[\sigma_p^q\left(X_{pn}+Y_{pn}\right)-\sigma_p^q\left(X_{pn}-Y_{pn}\right)\right] \quad (27)$$

$$X_{pn}=\frac{X-E\left[X\right]}{\sigma_p(X)} \quad (28)$$

$$cov_{p,q}(X,Y)=\frac{1}{4}\left[\sigma_p^q(X+Y)-\sigma_p^q(X-Y)\right] \quad (29)$$

$$Cov_{p,2}(X,Y)=c_p^{\frac{2}{p}}\cdot Cov(X,Y) \quad (30)$$

The classical expressions of the covariance and correlation are obtained for *p=q=2*. Other important particular case are for *p=q=1*, [*p=1, q=2*] and [*p=2, q=1*].

Generalized covariance and correlation can also be defined using a GMD based on the logarithm function or other functions. In the case when both random variables have a normal distribution, the generalized relations give similar results as the classical expressions, and little or any additional information can be obtained. The generalized relations can be useful when non-Gaussian random variable are used, or in the case of mixtures of Gaussians distributions.

The presence of outliers may be detected and affected correlation coefficients can be corrected choosing an adequate generalized correlation coefficient.

In the particular case of p=1 and p=2 the GMD are $\sigma_1(X)$ (31) and $\sigma_2(X)$ (32).

$$\sigma_1(X)=E\left[\left|X-E\left[X\right]\right|\right]=\sigma(X)\cdot\sqrt{\frac{2}{\pi}} \quad (31)$$

$$\sigma_2(X)=\left(E\left[\left|X-E\left[X\right]\right|^2\right]\right)^{\frac{1}{2}}=\sigma(X) \quad (32)$$

The correlation coefficients $corr_{1,1}(X,Y)$ and $corr_{2,1}(X,Y)$ are given by Eq. (33) and respectively by Eq. (34), and these two relations gives the same result because the constant c_1 is simplified (35).

$$corr_{1,1}(X,Y)=\frac{1}{4}\left[\sigma_1^1(X_{1n}+Y_{1n})-\sigma_1^1(X_{1n}-Y_{1n})\right] \quad (33)$$

$$corr_{2,1}(X,Y)=\frac{1}{4}\left[\sigma_2^1(X_{2n}+Y_{2n})-\sigma_2^1(X_{2n}-Y_{2n})\right] \quad (34)$$

$$\sigma_1^1(X_{1n}\pm Y_{1n})=\sigma_1^1\left(\frac{X}{\sqrt{\frac{2}{\pi}}\cdot\sigma(X)}\pm\frac{Y}{\sqrt{\frac{2}{\pi}}\cdot\sigma(Y)}\right) \quad (35) \\ =\sigma\left(\frac{X}{\sigma(X)}\pm\frac{Y}{\sigma(Y)}\right)$$

4. Pseudo-metrics functions

We define the statistical distance between two random variables *X* and *Y* with the Eq. (36) that can be rewritten as (37) where X_n and Y_n are the normalized random variables. This statistical distance is a pseudo metric function that satisfies the conditions

$$d(X,Y)=\frac{\sqrt{2}}{4}\cdot\left(\sigma\left[\frac{X}{\sigma(X)}+\frac{Y}{\sigma(Y)}\right]+\sigma\left[\frac{X}{\sigma(X)}-\frac{Y}{\sigma(Y)}\right]\right)- \\ -\frac{\sqrt{2}}{4}\cdot\left|\sigma\left[\frac{X}{\sigma(X)}+\frac{Y}{\sigma(Y)}\right]-\sigma\left[\frac{X}{\sigma(X)}-\frac{Y}{\sigma(Y)}\right]\right| \quad (36)$$

$$d(X,Y)=\frac{\sqrt{2}}{4}\cdot[\sigma(X_n+Y_n)+\sigma(X_n-Y_n)]- \\ -\frac{\sqrt{2}}{4}\cdot|\sigma(X_n+Y_n)-\sigma(X_n-Y_n)| \quad (37)$$

The first condition is fulfilled because in Eq. (37) the sum of two positive numbers is greater than their difference. The condition in the case when $d(X,Y)$ is a metric, is realized *if and only if* *X*=*Y*, but for random variables such a restriction is too strong. In this case, the condition is fulfilled, but $d(X,Y)=0$ also when the variables *X* and *Y* are fully correlated.

The condition is also fulfilled because $\sigma(X_n+Y_n)=\sigma(Y_n+X_n)$ and $\sigma(X_n-Y_n)=\sigma(Y_n-X_n)$ in Eq.(37). Condition can be proved observing that $d(X,Z)$ is maxim when *X* and *Y* are uncorrelated. $d(X,Y)+d(Y,Z)$ is minim when *Y* is correlated with *X* and *Z*, and in these conditions the Eqs. are fulfilled. The resulting inequality is rather simple and can be graphically resolved.

$$Y_n=a\cdot X_n+b\cdot Z_n \quad (44)$$

$$\overline{X_n\cdot Y_n}=a,\quad \overline{Y_n\cdot Z_n}=b \quad a^2+b^2=1 \quad (45)$$

$$\overline{X_n^{\,2}}=\overline{Y_n^{\,2}}=\overline{Z_n^{\,2}}=1 \quad (46)$$

A similar pseudo metric function can be defined using the MGD operator instead of the standard deviation (38). In the case when X and Y are normal distributed random variables the value of this distance is identical with (37). This can be proved using (35).

$$d_a(X,Y)=\frac{\sqrt{2}}{2}\cdot[D(X_{an}+Y_{an})+D(X_{an}-Y_{an})]- \\ -\frac{\sqrt{2}}{2}\cdot|D(X_{an}+Y_{an})-D(X_{an}-Y_{an})| \quad (38)$$

The values of the statistical distance for fully correlated or identical random variables is zero, for partially correlated variables its value is between zero and one, and for completely uncorrelated or independent variables the value is one.

5. Example

We consider three (pseudo) random variables x, y and z generated with MatLab:

x = random ('norm', 0, 1, nr_s, 1);
z = random('unif',-1, 1, nr_s,1) - x/3;
y = a*x + (1 - a)*z;

x has a normal distribution with zero mean and a standard deviation equal with 1, *z* is composed from a random variable with a uniform distribution in the interval (-1,1) from which is subtracted one third of *x*.

y is a linear combination of *x* and *z*, and this combination depends on the value of the parameter *a*. The samples number for each variable is nr_s=10^6.

The value of $d_a(x,z)$ is a constant, and it is represented in Fig. 2 with a red horizontal dotted line. On the abscise axis are represented the values of the parameter a. On ordinate are represented the values of $d(x,y)$ and $d(y,z)$ with a solid blue line and respectively with a dash line, the values of the $d_a(x,y)$ and $d_a(y,z)$ with a solid and dash blue line, the values of the correlation coefficients values $corr(x,y)$ and $corr(y,z)$ with a green and a dash green line.

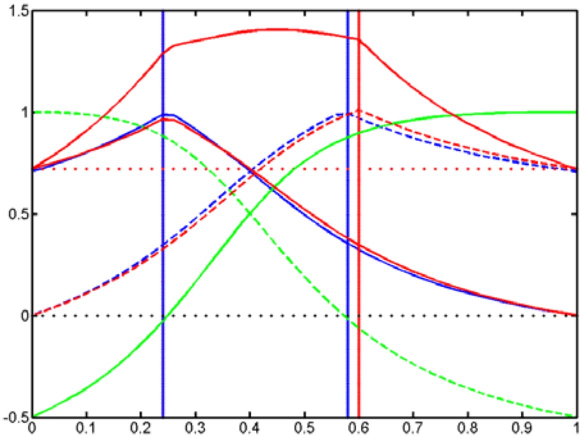

Fig. 2. The values of the statistical distance function of the parameter a.

When the value of the parameter a varies from zero to one the values of $d(x,y)$ and $d_a(x,y)$ to a maximum value of one and then decreases to zero. When the value of a=0.25 the variables x and y become independent and the statistical "distance" reaches the greatest value of one. In this point the correlation coefficient $corr(x,y)$=0. When a=1 the variables x and y become equals and the statistical "distance" is zero.

The graph of the sun of the two functions $d_a(x,y)$ and $d_a(y,z)$ is the red curve in the upper part of the Fig. 2.

The functions $d_a(y,z)$ and $d(y,z)$ varies similarly, when *a* =0 their values are zero because the variables *y* and *z* are identical. The functions reach their maximum value in different points. This difference is caused by the fact that in this point the correlation coefficient $corr(y,z)$ is zero, but the variables y and z are not independent, and such situations can be detected in this way.

If X_1 and X_2 are two independent random variables and X and Y are two linear combinations of these (39) then the correlation coefficient is zero for the values of *a₁,* *a₂,b₁*, and *b₂* that are the solutions of the equation (40). Not all the random variables X and Y that satisfy (40) are independent. Some of these situations can be detected using the two generalized correlation coefficients $corr_{1,1}(X,Y)$ and $corr_{2,2}(X,Y)$ that has different values.

$$X = a_1 \cdot X_1 + a_2 \cdot X_2, \quad Y = b_1 \cdot X_1 + b_2 \cdot X_2 \tag{39}$$

$$corr(X,Y) = a_1 \cdot b_1 \cdot \sigma^2(X_1) + a_2 \cdot b_2 \cdot \sigma^2(X_2) = 0 \tag{40}$$

The random variables *X* and *Y* that are solutions of the Eq. (40) for the same values of a_1, a_2, b_1, and b_2 may be independent or not if X_1 and X_2 has different distributions, and their standard deviations are the same. For example, if the independent random variable pair X_1 , X_2 follows the bivariate normal distribution then $X = X_1 + a \cdot X_2$ and $Y = X_1 - a \cdot X_2$ are independent random variables if $a = \frac{\sigma(X_1)}{\sigma(X_2)}$, but this might not be true if X_1 and X_2 has a different distribution. When both variables *x* and *y* are Gaussian then the two statistical distances $d_a(y,z)$, $d(y,z)$ gives the same values.

The statistical distance has been used for gaze guidance evaluation and for scan path measurements. The scan path is determinate mainly by the scene and the scene dynamics but it has also physiological components. Some of these physiological aspects has already been observed and studied, as for example, the gaze of tired subjects that is less dynamic. The scan path has also an important random component and some aspects are quite difficult to be observed and measured [5]. Processing many recordings have been observed that the statistical distance between the duration of a fixation and the amplitude of the next saccade is different for some subjects, and the fixation and the amplitude of the last saccade are less correlated.

6. Conclusions

The generalized covariance can be considered as an alternative to the classical covariance. It some cases it can be computed faster than the classical relation. Another advantage is that outliers less affect some particular relations of the generalized covariance than the classical covariance.

The generalized correlation coefficient has similar advantages as the generalized covariance. What is important to mention is that in this case is not necessary to know a conversion constant like in the case of the GMD and the generalized covariance. The values computed with some particular generalized correlation coefficient differ only by 1% from those computed with the Pearson's correlation coefficient.

The presence of outliers can be detected by computing the difference between different generalized correlation coefficients

The GMD the generalized covariance and correlation coefficient to be less sensitive to outliers have been tested on some experimental data with good results.

Acknowledgements

Our research has received funding from the European Commission through the GazeCom project under contract IST-C-033816 of the Sixth Framework Programme. All views expressed herein are those of the authors alone. The European Community is not liable for any use made of the information.

References

[1] C. Bradley, The Mathematical Gazette **69**, 12 (1985).
[2] S. K. Mishra, On Construction of Robust Composite Indices by Linear Aggregation, SSRN eLibrary, 2008.
[3] S. K. Mishra, Construction of Composite Indices in Presence of Outliers, SSRN eLibrary.
[4] S. K. Mishra, A Note on Positive Semi-Definiteness of Some Non-Pearsonian Correlation Matrices, SSRN eLibrary, 2009.
[5] E. Barth, et al., in Conference on Human Vision and Electronic Imaging XI, San Jose, CA, D570, 2006.

*Corresponding author: Dragos.Falie@cern.ch