INTERNATIONAL JOURNAL OF INTELLIGENT TECHNOLOGIES AND APPLIED STATISTICS
VOL.6, NO.3 (2013) PP.215–224, DOI: 10.6148/IJITAS.2013.0603.01

Towards a Localized Version of Pearson's Correlation Coefficient

Vladik Kreinovich[1*], Hung T. Nguyen[2,3] and Berlin Wu[4]

[1] *Department of Computer Sciences, University of Texas, El Paso, USA*
[2] *Department of Mathematical Sciences, New Mexico State University, New Mexico, USA*
[3] *Department of Economics, Chiang Mai University, Chiang Mai, Thailand*
[4] *Department of Mathematical Sciences, National Chengchi University, Taipei, Taiwan*

ABSTRACT

Pearson's correlation coeffcient is used to describe dependence between random variables X and Y. In some practical situations, however, we have strong correlation for some values X and/or Y and no correlation for other values of X and Y. To describe such a local dependence, we come up with a natural localized version of Pearson's correlation coefficient. We also study the properties of the newly defined localized coefficient.

Keywords: Pearson's correlation coefficient; Local correlation; Copulas

1. Formulation of the problem

1.1 Pearson's correlation coefficient: reminder

To describe relation between two random variables X and Y, Pearson's correlation coefficient r is often used; see, e.g., [3]. This coefficient is defined as

$$r[X,Y] \stackrel{\text{def}}{=} \frac{C[X,Y]}{\sigma[X]\cdot\sigma(Y)}, \tag{1}$$

where $C(X, Y) \stackrel{\text{def}}{=} E[(X - E[X]) \cdot (Y - E[Y])] = E[X \cdot Y] - E[X] \cdot E[Y]$ is the covariance, $E[X]$ means the mean, $\sigma[X] \stackrel{\text{def}}{=} \sqrt{V[X]}$ if the standard deviation, and $V[X] \stackrel{\text{def}}{=} E[(X - E[X])^2\,] = E[X^2] - (E[X])^2$ is the variance. Pearson's correlation coefficient ranges

* Corresponding author: vladik@utep.edu

between –1 and 1. When the variables X and Y are independent, then $r[X, Y] = 0$. When $r[X, Y] = 1$ or $r[X, Y] = -1$, this means that Y is a linear function of X (i.e., informally, that we have a perfect correlation).

1.2 Need for a local version of Pearson's correlation coefficient

Pearson's correlation coefficient provides a global description of the relation between the random variables X and Y. In some practical situations, there is a stronger correlation for some values of X and/or Y and a weaker correlation for other values of X and/or Y. To describe such local dependence, we need to come up with a local version of Pearson's correlation coefficient.

2. Towards a definition of a local version of Pearson's correlation coefficient

2.1 Motivation for the new definition

For given random variables X and Y and for given real numbers x and y, we want to describe the dependence between the variables X and Y limited to small neighborhood $(x - \varepsilon, x + \varepsilon) \times (y - \delta, y + \delta)$ of a point (x, y). This means, in effect, that instead of the pair of random variables (X, Y) corresponding to the original probability distribution, we consider a pair $(X_{x\pm\varepsilon,y\pm\delta}, Y_{x\pm\varepsilon,y\pm\delta})$ with the conditional probability distribution, under the condition that $(X, Y) \in (x - \varepsilon, x + \varepsilon) \times (y - \delta, y + \delta)$.

This conditional probability distribution can be described in the usual way: for every measurable set $S \subseteq \mathrm{IR}^2$, the corresponding probability

$$\mathrm{Prob}\left((X_{x\pm\varepsilon,y\pm\delta}, Y_{x\pm\varepsilon,y\pm\delta}) \in S\right)$$

is defined as

$$\mathrm{Prob}\left((X_{x\pm\varepsilon,y\pm\delta}, Y_{x\pm\varepsilon,y\pm\delta}) \in S\right) = \frac{\mathrm{Prob}\left((X,Y) \in S \cap (x-\varepsilon, x+\varepsilon) \times (y-\delta, y+\delta)\right)}{\mathrm{Prob}\left((X,Y) \in (x-\varepsilon, x+\varepsilon) \times (y-\delta, y+\delta)\right)}.$$

To describe the desired dependence, it is reasonable to consider the asymp- totic behavior of the correlation between $X_{x\pm\varepsilon,y\pm\delta}$ and $Y_{x\pm\varepsilon,y\pm\delta}$ when $\varepsilon \to 0$ and $\delta \to 0$. It turns out that for probability distributions with twice continuously differentiable probability density function $\rho(x, y)$, we can have an explicit expression for the Pearson's correlation coefficient $r[X_{x\pm\varepsilon,y\pm\delta}, Y_{x\pm\varepsilon,y\pm\delta}]$ in terms of $\rho(x, y)$:

Proposition 1 For probability distributions with twice continuously differentiable probability density functions $\rho(x, y)$, we have

$$r[X_{x\pm\varepsilon,y\pm\delta}, Y_{x\pm\varepsilon,y\pm\delta}] = \frac{1}{3}\cdot\varepsilon\cdot\delta\cdot\frac{\partial^2 \ln(\rho(x,y))}{\partial x\,\partial y} + o(\varepsilon\cdot\delta). \quad (2)$$

This asymptotic behavior is determined by a single parameter $\frac{\partial^2 \ln(\rho(x,y))}{\partial x\,\partial y}$

It is therefore reasonable to use this parameter as a local version of Pearson's correlation coefficient:

Definition 1 *Let (X, Y) be a random 2-D vector, and let (x, y) be a 2-D point. By the local correlation at a point (x, y), we mean the vvalue*

$$r_{x,y}[X,Y] \stackrel{\text{def}}{=} \frac{\partial^2 \ln(\rho(x,y))}{\partial x\,\partial y}. \quad (3)$$

Proof of Proposition Since the probability density function is twice diffierentiable, in the small vicinity of the point (x, y), we have

$$\rho(x+\Delta x, y+\Delta y) = c + c_x\cdot\Delta x + c_y\cdot\Delta y +$$
$$\frac{1}{2}\cdot c_{xx}\cdot(\Delta x)^2 + c_{xy}\cdot\Delta x\cdot\Delta y + \frac{1}{2}\cdot c_y y\cdot(\Delta y)^2 + o(\varepsilon^2,\delta^2), \quad (4)$$

where

$$c \stackrel{\text{def}}{=} \rho(x,y),\quad c_x \stackrel{\text{def}}{=} \frac{\partial\rho}{\partial x},\quad c_y \stackrel{\text{def}}{=} \frac{\partial\rho}{\partial y},$$
$$c_{xx} \stackrel{\text{def}}{=} \frac{\partial^2\rho}{\partial x^2},\quad c_x \stackrel{\text{def}}{=} \frac{\partial^2\rho}{\partial x\,\partial y},\quad c_{yy} \stackrel{\text{def}}{=} \frac{\partial^2\rho}{\partial y^2}. \quad (5)$$

The condition probability distribution is obtained by obtained by dividing the original one by

$$\mathrm{Prob}\,((X,Y)\in(x-\varepsilon,x+\varepsilon)\times(y-\delta,y+\delta)) =$$
$$\int_{-\varepsilon}^{\varepsilon}\int_{-\delta}^{\delta}\rho(x+\Delta x,y+\Delta y)\,d\Delta x\,d\Delta y = c\cdot(2\varepsilon)\cdot(2\delta)+o(\varepsilon,\delta) = 4c\cdot\varepsilon\cdot\delta+o(\varepsilon\delta). \quad (6)$$

Pearson's correlation coefficient does not change if we shift both X and Y by a constant, i.e., consider shifted random variables $\Delta X \stackrel{\text{def}}{=} X_{x\pm\varepsilon,y\pm\delta} - x$ and $\Delta Y \stackrel{\text{def}}{=} Y_{x\pm\varepsilon,y\pm\delta} - y$ instead of the original random variables $X_{x\pm\varepsilon,y\pm\delta}$ and $Y_{x\pm\varepsilon,y\pm\delta}$:

$$C[X_{x\pm\varepsilon,y\pm\delta}, Y_{x\pm\varepsilon,y\pm\delta}] = C[\Delta X, \Delta Y] = \frac{E[\Delta X \cdot \Delta Y] - E[\Delta X] \cdot E[\Delta Y]}{\sqrt{E[(\Delta X)^2] - (E[\Delta X])^2} \cdot \sqrt{E[(\Delta Y)^2] - (E[\Delta Y])^2}}. \tag{7}$$

Here,

$$E[\Delta X] = \frac{\int_{-\varepsilon}^{\varepsilon} \int_{-\delta}^{\delta} \Delta x \cdot \rho(x + \Delta x, y + \Delta y)\, d\Delta x\, d\Delta y}{\int_{-\varepsilon}^{\varepsilon} \int_{-\delta}^{\delta} \rho(x + \Delta x, y + \Delta y)\, d\Delta x\, d\Delta y}. \tag{8}$$

In this formula,

$$\Delta x \cdot \rho(x + \Delta x, y + \Delta y) = \Delta x \cdot (c + c_x \cdot \Delta x + c_y \cdot \Delta y + \frac{1}{2} \cdot c_{xx} \cdot (\Delta x)^2 + c_{xy} \cdot \Delta x \cdot \Delta y + \frac{1}{2} \cdot c_y y \cdot (\Delta y)^2 + o). \tag{9}$$

The integral, over a symmetric box, of any expression which is odd in Δx and/or Δy is equal to 0. Thus, the main term in this expression that leads to a non-zero integral is $c_x \cdot (\Delta x)^2$. For this term, the integral in the numerator of the formula (8) is equal to $c_x \cdot (2\delta) \cdot \left(\frac{\varepsilon^3}{3} - \frac{(-\varepsilon)^3}{3}\right) = c_x \cdot \delta \cdot \frac{4}{3} \cdot \varepsilon^3$. We already know that the denominator (6) of the formula (8) is equal to $4c \cdot \varepsilon \cdot \delta + o$, thus the formula (8) leads to:

$$E[\Delta X] = \frac{1}{3} \cdot \varepsilon^2 \cdot \frac{c_x}{c}. \tag{10}$$

Similarly,

$$E[\Delta Y] = \frac{1}{3} \cdot \delta^2 \cdot \frac{c_y}{c}. \tag{11}$$

For the expected value of $(\Delta X)^2$, we get

$$E[(\Delta X)^2] = \frac{\int_{-\varepsilon}^{\varepsilon}\int_{-\delta}^{\delta}(\Delta x)^2 \cdot \rho(x+\Delta x, y+\Delta y)\, d\Delta x\, d\Delta y}{\int_{-\varepsilon}^{\varepsilon}\int_{-\delta}^{\delta}\rho(x+\Delta x, y+\Delta y)\, d\Delta x\, d\Delta y} =$$

$$\frac{\int_{-\varepsilon}^{\varepsilon}\int_{-\delta}^{\delta}(c\cdot(\Delta x)^2 + o)\, d\Delta x\, d\Delta y}{\int_{-\varepsilon}^{\varepsilon}\int_{-\delta}^{\delta}(c+o)\, d\Delta x\, d\Delta y} = \frac{c\cdot 2\cdot\delta\cdot\left(\frac{\varepsilon^3}{3} - \frac{(-\varepsilon)^3}{3}\right) + o}{4c\cdot\varepsilon\cdot\delta + o} = \frac{1}{3}\cdot\varepsilon^2. \tag{12}$$

Here, $E[(\Delta X)^2] \sim \varepsilon^2$ and, due to (10), $(E[\Delta X])^2 \sim \varepsilon^4 = o(\varepsilon^2)$. Thus,

$$E[(\Delta X)^2] - (E[\Delta X])^2 = E[(\Delta X)^2] + o = \frac{1}{3}\cdot\varepsilon^2 + o. \tag{13}$$

Similarly,

$$E[(\Delta Y)^2] - (E[\Delta Y])^2 = \frac{1}{3}\cdot\varepsilon^2 + o, \tag{14}$$

so the denominator of the expression (7) is equal to

$$\sqrt{E[(\Delta X)^2] - (E[\Delta X])^2}\cdot\sqrt{E[(\Delta Y)^2] - (E[\Delta Y])^2} = \frac{1}{3}\cdot\varepsilon\cdot\delta + o. \tag{15}$$

For the numerator of the formula (7), we get

$$E[\Delta X\cdot\Delta Y] = \frac{\int_{-\varepsilon}^{\varepsilon}\int_{-\delta}^{\delta}\Delta x\cdot\Delta y\cdot\rho(x+\Delta x, y+\Delta y)\, d\Delta x\, d\Delta y}{\int_{-\varepsilon}^{\varepsilon}\int_{-\delta}^{\delta}\rho(x+\Delta x, y+\Delta y)\, d\Delta x\, d\Delta y}. \tag{16}$$

In this formula,

$$\begin{gathered}\Delta x\cdot\Delta y\cdot\rho(x+\Delta x, y+\Delta y) = \\ \Delta x\cdot\Delta y\cdot(c + c_x\cdot\Delta x + c_y\cdot\Delta y + \frac{1}{2}\cdot c_{xx}\cdot(\Delta x)^2 + c_{xy}\cdot\Delta x\cdot\Delta y + \frac{1}{2}\cdot c_y y\cdot(\Delta y)^2 + o).\end{gathered} \tag{17}$$

The only term here which is not odd in Δx or in Δy is the term $c_{xy}\cdot(\Delta x)^2\cdot(\Delta y)^2$, for which the integral in the numerator of (16) is equal to $c_{xy}\cdot\frac{2}{3}\cdot\varepsilon^3\cdot\frac{2}{3}\cdot\delta^3 + o$. Since the denominator (6) of the expression (16) is equal to $4c\cdot\varepsilon\cdot\delta + o$, thus

$$E[\Delta X \cdot \Delta Y] = \frac{1}{9} \cdot \frac{c_{xy}}{c} \cdot \varepsilon^2 \cdot \delta^2 + o. \tag{18}$$

From (10), (11) and (18), we conclude that

$$E[\Delta X \cdot \Delta Y] - E[\Delta X] \cdot E[\Delta Y] = \frac{1}{9} \cdot \varepsilon^2 \cdot \delta^2 \cdot \left(\frac{c_{xy}}{c} - \frac{c_x \cdot c_y}{c^2}\right) + o. \tag{19}$$

From (19) and (15), we can now conclude that the ratio (7) has the form

$$C[X_{x\pm\varepsilon,y\pm\delta}, Y_{x\pm\varepsilon,y\pm\delta}] = C[\Delta X, \Delta Y] = \frac{1}{3} \cdot \varepsilon \cdot \delta \cdot \left(\frac{c_{xy}}{c} - \frac{c_x \cdot c_y}{c^2}\right) + o. \tag{20}$$

Substituting the expressions (5) for c, c_x c_y, and c_{xy} in terms of the probability density $\rho(x, y)$ and its derivative, we can easily check that the expression

$$\frac{c_{xy}}{c} - \frac{c_x \cdot c_y}{c^2}$$

is indeed equal to the derivative (3). The proposition is proven.

3. What is the meaning of the new definition for a normal distribution

To better understand the meaning of our newly defined term, let us compute its value for a normal distribution, for which

$$\rho(x, y) =$$

$$\text{const} \cdot \exp\left(-\frac{a_{xx} \cdot (x - x_0)^2 + 2a_{xy} \cdot (x - x_0) \cdot (y - y_0) + a_{yy} \cdot (y - y_0)^2}{2}\right),$$

where the matrix

$$a = \begin{pmatrix} a_{xx} & a_{xy} \\ a_{xy} & a_{yy} \end{pmatrix}$$

is the inverse matrix to the covariance matrix

$$C = \begin{pmatrix} V[X] & C[X,Y] \\ C[X,Y] & V[Y] \end{pmatrix}. \tag{21}$$

For this distribution,

$$\ln(\rho(x,y)) = -\frac{1}{2}\cdot(a_{xx}\cdot(x-x_0)^2+2a_{xy}\cdot(x-x_0)\cdot(y-y_0)+a_{yy}\cdot(y-y_0)^2)), \tag{22}$$

and thus,

$$r_{x,y}[X,Y] = \frac{\partial^2 \ln(\rho(x,y))}{\partial x\,\partial y} = -a_{xy}. \tag{23}$$

If we use an explicit formula for the elements of the inverse 2×2 matrix to describe a_{xy} in terms of the element of the covariance matrix (21), we get the following expression:

$$r_{x,y}[X,Y] = \frac{C[X,Y]}{V[X]\cdot V[Y]-(C[X,Y])^2}. \tag{24}$$

Substituting $V[X] = (\sigma[X])^2$, $V[Y] = (\sigma[Y])^2$, and

$$C[X, Y] = \sigma[X] \cdot \sigma[Y] \cdot r[X, Y]$$

into the formula (24), we conclude that

$$r_{x,y}[X,Y] = \frac{r_{x,y}[X,Y]}{1-(r_{x,y}[X,Y])^2}\cdot\frac{1}{\sigma[X]\cdot\sigma[Y]}. \tag{25}$$

4. Criterion for independence

It turns out that the new localized version of Pearson's correlation coefficient provides a natural criterion for independence.

Proposition 2 For a random 2-D vector (X, Y) with a twice continuously differentiable probability density function $\rho(x, y)$, the following two conditions are equivalent to each other:

- X and Y are independent;
- $r_{x,y}[X, Y] = 0$ for all x and y.

Proof If X and Y are independent, then $\rho(x, y) = \rho_X(x) \cdot \rho_Y(y)$, where $\rho_X(x)$ and $\rho_Y(y)$ are probability densities corresponding to the marginal distributions. Thus, $\ln(\rho(x, y)) = \ln(\rho_X(x)) + \ln(\rho_Y(y))$ and therefore, for every x and y, we have have. $r_{x,y}[X,Y] = \dfrac{\partial^2 \ln(\rho(x,y))}{\partial x\, \partial y} = 0.$

Vice versa, let us assume that $r_{x,y}[X, Y] = \dfrac{\partial^2 \ln(\rho(x,y))}{\partial x\, \partial y} = 0$ for all x and y. The fact that the x-partial derivative of the auxiliary function $\dfrac{\partial \ln(\rho(x,y))}{\partial y}$ is equal to 0 means that this auxiliary function does not depend on x, i.e., that it depends only on y:

$$\frac{\partial \ln(\rho(x,y))}{\partial y} = f_1(y) \tag{26}$$

for some function $f_1(y)$. Integrating over y, we get

$$\ln(\rho(x,y)) = f_2(y) + f_3(x), \tag{27}$$

where $f_2(y) \stackrel{\text{def}}{=} \int_0^y f_1(t)\, dt$ is an integral of the function $f_1(y)$, and $f_3(x)$ is a constant of integration, constant which may depend on x. For $\rho(x, y) = \exp(\ln(\rho(x, y))$, we thus conclude that

$$\rho(x, y) = \exp(F_3(x)) \cdot \exp(F_2(y)) \tag{28}$$

where $F_3(x) \stackrel{\text{def}}{=} (f_3(x))$ and $F_2(y) \stackrel{\text{def}}{=} (f_2(y))$. Since $\rho(x, y) \geq 0$ for all x and y, we can conclude that

$$\rho(x,y) = |F_3(x)| \cdot |F_2(y)|, \tag{29}$$

with $|F_3(x)| \geq 0$ and $|F_2(x)| \geq 0$. By normalizing the functions of x and y, i.e., by taking $\rho_X(x) \dfrac{|F_3(x)|}{\int_{-\infty}^{\infty} |F_3(t)|\, dt}$ and $\rho_Y(y) \stackrel{\text{def}}{=} \dfrac{|F_2(y)|}{\int_{-\infty}^{\infty} |F_2(t)|\, dt}$ we conclude that $\rho(x, y) = \rho_X(x) \cdot \rho_Y(y)$, i.e., that X and Y are indeed independent. The proposition is proven.

5 . Relation to copulas

A probability distribution with a probability distribution function $F(x, y) \stackrel{\text{def}}{=}$ Prob $(X \le x \;\&\; Y \le y)$ can be described as

$$F(x, y) = C(F_X(x), F_Y(y)), \tag{30}$$

where $F_X(x) \stackrel{\text{def}}{=}$ Prob $(X \le x)$ and $F_Y(y) \stackrel{\text{def}}{=}$ Prob $(Y \le y)$ are marginal distri-butions, and a function $C(a, b) \stackrel{\text{def}}{=} F(F_X^{-1}(a), F_Y^{-1}(b))$ is known as a *copula*; see, e.g., [1-2].

A copula can also be viewed as a probability distribution function for a 2-D random vector (A, B), with a probability density function $c(a, b) \stackrel{\text{def}}{=} \dfrac{\partial^2 C(a,b)}{\partial a\, \partial b}$. The probability density function $\rho(x, y)$ of the original distribution can be described, in terms of the cumulative distribution function $F(x, y)$, as $\rho(x, y) = \dfrac{\partial^2 F(x,y)}{\partial x\, \partial y}$ Substituting the expression (30) into this formula and using the chain rule, we conclude that

$$\rho(x, y) = c(F_X(x), F_Y(y)) \cdot \rho_X(x) \cdot \rho_Y(y), \tag{31}$$

where $\rho_X(x)$ and $\rho_Y(y)$ are the probability densities corresponding to the marginal distributions.

For the copula's random vector (A, B), we can also define the local Pearson's correlation coefficient:

$$r_{a,b}[A, B] = \frac{\partial^2 \ln(c(a, b))}{\partial a\, \partial b}. \tag{32}$$

It turns out that the above localized version of Pearson's correlation coefficient can be naturally reformulated in terms of the copula and marginal distributions; namely, the relation is the same as the relation (31) for probability densities:

Proposition 3 Let (X, Y) be a random 2-D vector with marginal distributions $F_X(x)$ and $F_Y(y)$, and let (A, B) be the corresponding copula distribution. Then,

$$r_{x,y}[X, Y] = r_{F_X(x), F_Y(y)}[A, B] \cdot \rho_X(x) \cdot \rho_Y(y). \tag{33}$$

Proof By taking logarithms of both sides of the formula (31), we get

$$\ln(\rho(x, y)) = \ln(c(F_X(x), F_Y(y)) + \ln(\rho_X(x)) + \ln(\rho_Y(y)). \tag{34}$$

Differentiating both sides of this formula with respect to x and y, we get the desired expression (33). The proposition is proven.

References

[1] P. Jaworski, F. Durante, W. K. Härdle, and T. Ruchlik (Eds.) (2010). *Copula Theory and Its Applications*, Springer Verlag, Berlin, Germany.

[2] R. B. Nelsen (1999). *An Introduction to Copulas*, Springer Verlag, New York.

[3] D. J. Sheskin (2011). *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida.