How to Detect Linear Dependence on the Copula Level?

Vladik Kreinovich[1], Hung T. Nguyen[2,3], and Songsak Sriboonchitta[3]

[1] Department of Computer Science, University of Texas at El Paso
500 W. University, El Paso, TX 79968, USA, vladik@utep.edu
[2] Department of Mathematical Sciences, New Mexico State University
Las Cruces, New Mexico 88003, USA, hunguyen@nmsu.edu
[3] Department of Economics, Chiang Mai University
Chiang Mai, Thailand, songsak@econ.chiangmai.ac.th

Abstract. In many practical situations, the dependence between the quantities is linear or approximately linear. Knowing that the dependence is linear simplifies computations; so, is is desirable to detect linear dependencies. If we know the joint probability distribution, we can detect linear dependence by computing Pearson's correlation coefficient. In practice, we often have a copula instead of a full distribution; in this case, we face a problem of detecting linear dependence based on the copula. Also, distributions are often heavy-tailed, with infinite variances, in which case Pearson's formulas cannot be applied. In this paper, we show how to modify Pearson's formula so that it can be applied to copulas and to heavy-tailed distributions.

1 Detecting Linear Dependence: Formulation of the Problem

Locally, linear dependencies are ubiquitous. Dependencies between quantities are often described by smooth (even analytical) functions $y = f(x_1, \ldots, x_n)$. An analytical function can be expanded in Taylor series around each point $x^{(0)} = (x_1^{(0)}, \ldots, x_n^{(0)})$:

$$y = f(x^{(0)}) + \sum_{i=1}^{n} c_i \cdot (x_i - x_i^{(0)}) + \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} \cdot (x_i - x_i^{(0)}) \cdot (x_j - x_j^{(0)}) + \ldots \quad (1)$$

For values x_i close to $x_i^{(0)}$, we can safely ignore terms which are quadratic in $x_i - x_i^{(0)}$ (or of higher order), and thus, approximate the dependence by a linear function $y \approx f(x^{(0)}) + \sum\limits_{i=1}^{n} c_i \cdot (x_i - x_i^{(0)})$.

Linear dependencies are often global. In many practical situations, linear dependencies extend beyond local, they hold even for situations in which differences $x_i - x_i^{(0)}$ are reasonably large.

It is important to know if we have a linear dependence. Linear dependencies make computations easier. For example, there are efficient algorithms for solving systems of linear equations, while a solution to the system of non-linear equations is, in general, NP-hard; see, e.g., [12].

An exact linear dependence is easy to detect. Let us first consider the ideal case, when estimation and measurement errors can be safely ignored, and the dependence is exactly linear. In this case, if we have K situations in which we measured all the values x_i and y, then, based on the corresponding values $(x_1^{(k)}, \ldots, x_n^{(k)}, y^{(k)})$, $k = 1, 2, \ldots, K$, we can check the dependence is linear by checking whether the corresponding system of linear equations with unknowns c_i has a solution:

$$y^{(k)} = f(x^{(0)}) + \sum_{i=1}^{n} c_i \cdot \left(x_i^{(k)} - x_i^{(0)}\right), \quad k = 1, \ldots, K. \tag{2}$$

As we have mentioned, there exist efficient algorithms for checking solvability of such a linear system.

How the presence of an approximate linear dependence is detected now. Since linear dependencies make computations easier, it is desirable to detect them even when we only have an approximate linear dependence: e.g., due to measurement or approximation errors, or due to actual non-linear terms in the dependence, or due to the fact that the value of the quantity y is only approximately determined by the values $x_1, \ldots, x_n$.

In the case of the exact linear dependence, possible values of the tuple $(x_1, \ldots, x_n, y)$ form a linear surface $y = f(x^{(0)}) + \sum\limits_{i=1}^{n} c_i \cdot (x_i - x_i^{(0)})$. When we observe the frequency with which different tuples occur, we get a probability distribution on this surface.

In the case of an approximate linear dependence, tuples can deviate from the surface corresponding to the exact linear equation. In this case, the probability distribution is no longer limited to this surface. Instead, we have a probability distribution on the $(n + 1)$-dimensional space. Let $\rho(x_1, \ldots, x_n, y)$ denote the probability density of this probability distribution.

In traditional statistics, in the simplest case $n = 1$, the linearity of the corresponding dependence can be gauged by computing the Pearson's correlation coefficient (see, e.g., [25]):

$$r = \frac{C_{XY}}{\sigma_X \cdot \sigma_Y}, \tag{3}$$

where

$$C_{XY} \stackrel{\text{def}}{=} E[(X - E[X]) \cdot (Y - E[Y])] = E[X \cdot Y] - E[X] \cdot E[Y] =$$

$$\int x \cdot y \cdot \rho(x, y)\, dxdy - E[X] \cdot E[Y], \tag{4}$$

$$E[X] \stackrel{\text{def}}{=} \int x \cdot \rho(x,y)\,dxdy, \quad E[Y] \stackrel{\text{def}}{=} \int y \cdot \rho(x,y)\,dx, \tag{5}$$

$$\sigma_X \stackrel{\text{def}}{=} \sqrt{V_X}, \quad \sigma_Y \stackrel{\text{def}}{=} \sqrt{V_Y}, \tag{6}$$

$$V_X \stackrel{\text{def}}{=} E[(X - E[X])^2] = E[X^2] - (E[X])^2 =$$

$$\int x^2 \cdot \rho(x,y)\,dx\,dy - \left(\int x \cdot \rho(x,y)\,dx\,dy\right)^2, \tag{7}$$

$$V_Y \stackrel{\text{def}}{=} E[(Y - E[Y])^2] = E[Y^2] - (E[Y])^2 =$$

$$\int y^2 \cdot \rho(x,y)\,dx\,dy - \left(\int y \cdot \rho(x,y)\,dx\,dy\right)^2. \tag{8}$$

In the case of an exact linear dependence $y = c_0 + c_1 \cdot x$, this coefficient is equal to 1 if $c_1 > 0$ and to -1 if $c_1 < 0$. In general, values $r \neq 0$ indicate that there is an approximate linear dependence – and the closer $|r|$ to 1, the closer is to the actual dependence to a linear one.

Limitations of the existing techniques for detecting linear dependence. There are two major limitations of this technique:

- first, often, instead of the full joint probability distribution $\rho(x,y)$, we only know the corresponding *copula* (see below);
- second, the above formula only works when the variances are finite; in many practical situations, we have *heavy-tailed* distributions, in which variances are infinite.

What we do in this paper. In this paper, we show how to detect a linear dependence based on the copula, and we also show what to do when we have a heavy-tailed distribution.

2 Detecting Linear Dependence Based on a Copula: Formulation of the Problem

What is a copula: a brief reminder. Before we start analyzing how to detect linear dependence based on a copula, let us recall what is a copula.

In the general case, a distribution of a random variable X can be described by the cumulative distribution function $F_X(x) \stackrel{\text{def}}{=} \text{Prob}(X \leq x)$, and a joint distribution of two variables X and Y can be described by the cumulative distribution function $F(x,y) \stackrel{\text{def}}{=} \text{Prob}(X \leq x \,\&\, Y \leq y)$.

A problem with this description is that it depends on the units in which we describe x and y. For example, if we use meters instead of feet to describe x, or if we use a logarithmic scale of decibels instead of a linear scale of energy to describe noise, we get different cumulative distribution functions $F(x,y)$.

It is desirable to describe the dependence between x and y in a way which is independent on the units for measuring x and y. Such a description is known as a copula. The main idea behind a copula is that, once we know the probability distribution, we no longer need to use any artificial units to describe each of the quantities x and y:

- to describe the value of x, we can use the probability $F_X(x) = \text{Prob}(X \le x)$; and
- to describe the value of y, we can use the probability $F_Y(y) = \text{Prob}(Y \le y)$.

Thus, instead of asking for a value $F(x,y) = \text{Prob}(X \le x \,\&\, Y \le y)$ corresponding to given real numbers x and y, we can ask for a value $C(a,b)$ of this probability corresponding to given probabilities $a = F_X(x)$ and $b = F_Y(y)$.

Formally, the copula is defined as a function $C(a,b)$ for which $a = F_X(x)$ and $b = F_Y(y)$ imply that $F(x,y) = c(a,b)$, i.e., equivalently, as a function for which $F(x,y) = C(F_X(x), F_Y(y))$ for all x and y.

Copulas are useful. Copulas have been successfully used to describe dependencies in many application areas, including econometrics; see, e.g., [10, 20, 21].

Formulation of the problem. We need to be able to detect linear dependence between the quantities x and y based on the copula $C(a,b)$ that describes their dependence.

3 Detecting Linear Dependence Based on a Copula: Main Idea and the Resutling Definition

Main idea. As we have mentioned earlier, a copula describes two quantities x and y modulo arbitrary (generally, non-linear) monotonic re-scalings $X \to X' = A(X)$ and $Y \to Y' = B(Y)$. Even if the dependence is exactly linear (and $r = 1$ or $r = -1$) for some choice of scales, it stops being linear after a non-linear re-scaling.

What we want to check if whether the dependence is linear for *some* possible re-scaling. For different re-scalings, we have different values of the Pearson's correlation coefficient. The possibility to have $r = 1$ for *one* of the re-scalings means that the *maximum* L^+ of r over all such re-scaling is equal to 1. Thus, we can use this maximum to gauge to what a dependence described by a given copula is, in appropriate scales, described by an increasing linear function.

Similarly, the possibility to have $r = -1$ for *one* of the re-scalings means that the *minimum* L^- of r over all such re-scaling is equal to -1. Thus, we can use this minimum to gauge to what a dependence described by a given copula is, in appropriate scales, described by an decreasing linear function.

In general, we can use the pair (L^+, L^-) to detect a linear dependence based on the copula:

- the values $L^+ = 1$ or $L^- = -1$ means that (in appropriate scales) we have a perfect linear dependence;

– in general, In general, values $L^+ > 0$ or $L^- < 0$ indicate that there is an approximate linear dependence – and the closer $|L^+|$ or $|L^-|$ to 1, the closer is to the actual dependence to a linear one.

Resulting definition. Let us select any two variables X and Y with a given copula. All other variables with this same copula can be obtained from X and Y by applying appropriate non-linear transformations $X' = A(X)$ and $Y' = B(Y)$. The desired measures of linearity L can therefore be described as

$$L^+ = \max_{A(x),B(y)} r(A(X), B(Y)), \quad L^- = \min_{A(x),B(y)} r(A(X), B(Y)), \tag{9}$$

where maximum and minimum are taken over all possible non-decreasing functions $A(x)$ and $B(y)$ and $r(A(X), B(Y))$ is the Pearson's correlation coefficient relating the random variables $A(X)$ and $B(Y)$.

By definition of Pearson's correlation coefficient r, we conclude that

$$L^+ = \max_{A(x),B(y)} L(A, B); \quad L^+ = \min_{A(x),B(y)} L(A, B), \tag{10}$$

$$L(A, B) \stackrel{\text{def}}{=} \frac{C(A, B)}{\sigma(A) \cdot \sigma(B)}, \tag{11}$$

$$C(A, B) = E[(A(X) \cdot B(Y))] - E[A(X)] \cdot E[B(Y)] =$$

$$\int A(x) \cdot b(y) \cdot \rho(x, y)\, dx\, dy -$$

$$\left(\int A(x) \cdot \rho(x, y)\, dx\, dy\right) \cdot \left(\int B(y) \cdot \rho(x, y)\, dx\, dy\right), \tag{12}$$

$$\sigma(A) \stackrel{\text{def}}{=} \sqrt{V(A)}, \quad \sigma(B) \stackrel{\text{def}}{=} \sqrt{V(B)}, \tag{13}$$

$$V(A) \stackrel{\text{def}}{=} E[A^2(X)] - (E[A(X)])^2 =$$

$$\int A^2(x) \cdot \rho(x, y)\, dx\, dy - \left(\int A(x) \cdot \rho(x, y)\, dx\, dy\right)^2, \tag{14}$$

$$V(B) \stackrel{\text{def}}{=} E[B^2(X)] - (E[B(X)])^2 =$$

$$\int B^2(y) \cdot \rho(x, y)\, dx\, dy - \left(\int B(y) \cdot \rho(x, y)\, dx\, dy\right)^2. \tag{15}$$

These values depend only on the copula. The above definitions do not change if we re-scale each of the two variables. Thus, we conclude that these quantities depend only on the copula. To confirm this, let us show how the above definitions (10) can be explicitly reformulate in copula terms.

An explicit expression for L^+ and L^- in terms of the copula. The joint distribution of two random variables depends on the copula $C(a, b)$, and on the two marginal distributions $F_X(x)$ and $F_Y(y)$.

It is known that a general probability distribution for X can be described $X = A(X_0)$ for an appropriate function $A(x)$: namely, as $A(x)$, we can take an inverse function to $F_X(x)$. (This is one of the mostly used way to simulate a general probability distribution based on the uniform distribution – which is included in most programming languages.) Similarly, a general probability distribution for Y can be described $Y = B(Y_0)$ for an appropriate function $B(y)$. In this case, for each variable, $F_{X_0}(x) = \text{Prob}(X_0 \le x) = x$ and $F_{Y_0}(y) = \text{Prob}(Y_0 \le y) = y$ and thus, the joint cumulative distribution function has the form $F(x, y) = C(x, y)$ and thus, $\rho(x, y) = \dfrac{\partial^2 C(x, y)}{\partial x \partial y}$. For this probability density, the above formulas (10)–(15) become expressed solely in terms of a copula.

4 How to Compute the Corresponding Values L^+ and L^-

Analysis of the problem. According to calculus, one way to find minimum and maximum of an expression is to equate the derivative to 0. In our case, we need to situations when the unknowns are two functions $A(x)$ and $B(y)$, the rules for corresponding differentiation are described in variational calculus; see, e.g., [8].

Here, $\sigma(B)$ does not depend on $A(x)$, so, by using the usual rules of differentiating the ratio, we get:

$$\frac{\delta}{\delta A(x)} L(A, B) = \frac{1}{\sigma(B)} \cdot \frac{\delta}{\delta A(x)} \left(\frac{C(A, B)}{\sigma(A)} \right) =$$

$$\frac{1}{\sigma(B)} \cdot \frac{\delta}{\delta A(x)} \cdot \frac{\dfrac{\delta C(A, B)}{\delta A(x)} \cdot \sigma(A) - C(A, B) \cdot \dfrac{\delta \sigma(A)}{\delta A(x)}}{\sigma^2(A)}. \tag{16}$$

Thus, the derivative is equal to 0 if

$$\frac{\delta C(A, B)}{\delta A(x)} \cdot \sigma(A) - C(A, B) \cdot \frac{\delta \sigma(A)}{\delta A(x)} = 0. \tag{17}$$

Since $\sigma(A) = \sqrt{V(A)}$, the chain rule for differentiation implies that

$$\frac{\delta \sigma(A)}{\delta A(x)} = \frac{1}{2\sigma(A)} \cdot \frac{\delta V(A)}{\delta A(x)}. \tag{18}$$

For $V(A) = \int A^2(x) \cdot \rho(x, y)\, dx\, dy - \left(\int A(x) \cdot \rho(x, y)\, dx\, dy \right)^2$, we get

$$\frac{\delta V(A)}{\delta A(x)} = 2A(x) \cdot \int \rho(x, y)\, dy - 2E[A(X)] \cdot \int \rho(x, y)\, dy. \tag{19}$$

Similarly, for

$$C(A, B) = \int A(x) \cdot B(y) \cdot \rho(x, y)\, dx\, dy -$$

$$\left(\int A(x)\cdot\rho(x,y)\,dx\,dy\right)\cdot\left(\int B(y))\cdot\rho(x,y)\,dx\,dy\right), \tag{20}$$

we get

$$\frac{\delta C(A,B)}{\delta A(x)}=\int B(y)\cdot\rho(x,y)\,dx\,dy-E[B(Y)]\cdot\int\rho(x,y)\,dy. \tag{21}$$

Thus, the above equation (17) takes the form

$$c_1\cdot\int B(y)\cdot\rho(x,y)\,dx\,dy+c_2\cdot A(x)\cdot\int\rho(x,y)\,dy+c_3\cdot\int\rho(x,y)\,dy=0 \tag{22}$$

for some constants c_i. From this equation, we can determine $A(x)$ as

$$A(x)=a_1+a_2\cdot E[B(Y)\,|\,X=x], \tag{23}$$

where a_i are appropriate constants, and the conditional expected value

$$E[B(Y)\,|\,X=x] \tag{24}$$

has the form

$$E[B(Y)\,|\,X=x]=\frac{\displaystyle\int B(y)\cdot\rho(x,y)\,dx\,dy}{\displaystyle\int\rho(x,y)\,dx\,dy}. \tag{25}$$

By differentiating with respect to $B(y)$, we get a similar equation

$$B(y)=b_1+b_2\cdot E[A(X)\,|\,Y=y], \tag{26}$$

for appropriate constants b_1 and b_2.

These expressions depend on constants a_i and b_j which need to be determined. To make the expressions easier, we can take into account that the correlation coefficient does not change if we apply a linear transformation to the variables. Thus, instead of the functions $A(x)$ and $B(y)$, we can use arbitrary linear re-scalings $a+a'\cdot A(x)$ and $b+b'\cdot B(y)$. We can use this ambiguity to normalize the functions $A(x)$ and $B(y)$, e.g., by setting $A(0)=B(0)=0$ and $A(1)=B(1)=1$. By applying these conditions to the above formula for $B(y)$, we conclude that

$$B(0)=0=b_1+b_2\cdot E[A(X)\,|\,Y=0], \tag{27}$$

$$B(1)=1=b_1+b_2\cdot E[A(X)\,|\,Y=1]. \tag{28}$$

Substracting the first equation from the second one, we get

$$1=b_2\cdot(E[A(X)\,|\,Y=1]-E[A(X)\,|\,Y=0]), \tag{29}$$

hence

$$b_2=\frac{1}{E[A(X)\,|\,Y=1]-E[A(X)\,|\,Y=0]}. \tag{30}$$

From the equation (27) for $B(0)$, we can now conclude that

$$b_1 = -\frac{E[A(X)\,|\,Y=0]}{E[A(X)\,|\,Y=1] - E[A(X)\,|\,Y=0]}. \tag{31}$$

Substituting the expressions for b_1 and b_2 into the formula (26) for $B(y)$, we thus conclude that

$$B(y) = \frac{E[A(X)\,|\,Y=y] - E[A(X)\,|\,Y=0]}{E[A(X)\,|\,Y=1] - E[A(X)\,|\,Y=0]}. \tag{32}$$

Similarly, we get

$$A(x) = \frac{E[B(Y))\,|\,X=x] - E[B(Y)\,|\,X=0]}{E[B(Y)\,|\,X=1] - E[B(Y)\,|\,X=0]}. \tag{33}$$

Resulting algorithm. Formulas (32) and (33) prompts the following natural iterative algorithm. We start with arbitrary initial functions $A^{0)}(x)$ and $B^{0)}(y)$, e.g., with functions $A^{(0)}(x) = x$ and $B^{(0)}(y) = y$. Then, on each iteration, once we know the values $A^{(k)}(x)$ and $B^{(k)}(y)$, we compute the values corresponding to the next iteration as follows:

$$A^{(k+1)}(x) = \frac{E[B^{(k)}(Y))\,|\,X=x] - E[B^{(k)}(Y)\,|\,X=0]}{E[B^{(k)}(Y)\,|\,X=1] - E[B^{(k)}(Y)\,|\,X=0]}, \tag{34}$$

$$B^{(k+1)}(y) = \frac{E[A^{(k)}(X)\,|\,Y=y] - E[A^{(k)}(X)\,|\,Y=0]}{E[A^{(k)}(X)\,|\,Y=1] - E[A^{(k)}(X)\,|\,Y=0]}. \tag{35}$$

We stop when the new functions $A^{(k+1)}(x)$ and $B^{(k+1)}(y)$ are close to functions $A^{(k)}(x)$ and $B^{(k)}(y)$ from the previous iteration: e.g., when the differences do not exceed some threshold ε:

$$|A^{(k+1)}(x) - A^{(k+1)}(x)| \le \varepsilon; \quad |B^{(k+1)}(y) - B^{(k+1)}(y)| \le \varepsilon. \tag{36}$$

We then take $A^{(k+1)}(x)$ and $B^{(k+1)}(y)$ as the desired functions $A(x)$ and $B(y)$. Based on these finctions, we use the formula (11) to compute the desired value L^+.

Example. To make sure that this algorithm makes sense, let us analyze what happens when we apply this algorithm to the standard case of two jointly distributed correlated Gaussian variables.

Let us start with the simplest initial functions $A^{(0)}(x) = x$ and $B^{(0)}(y) = y$. For these functions, the formulas (34) and (35) for computing the next iteration $A^{(1)}(x)$ and $B^{(1)}(y)$ take the form

$$A^{(1)}(x) = \frac{E[Y\,|\,X=x] - E[Y\,|\,X=0]}{E[Y\,|\,X=1] - E[Y\,|\,X=0]}, \tag{37}$$

$$B^{(1)}(y) = \frac{E[X\,|\,Y=y] - E[X\,|\,Y=0]}{E[X\,|\,Y=1] - E[X\,|\,Y=0]}. \tag{38}$$

It is know that when variables X and Y have a Gaussian joint distribution, then $E[Y \mid X = x]$ is a linear function of x, i.e.,

$$E[Y \mid X = x] = c_0 + c_1 \cdot x \tag{39}$$

for some constant c_0 and c_1. Substituting this expression (30) into the formula (37), we get

$$A^{(1)}(x) = \frac{(c_0 + c_1 \cdot x) - (c_0 + c_1 \cdot 0)}{(c_0 + c_1 \cdot 1) - (c_0 + c_1 \cdot 0)} = \frac{c_1 \cdot x}{c_1} = x. \tag{40}$$

Similarly, we get $B^{(1)}(y) = y$.

Here, we have $A^{(1)}(x) = A^{(0)}(x)$ and $B^{(1)}(y) = B^{(0)}(y)$ for all x ad y, so we stop iterations, and take $A(x) = A^{(1)}(x) = x$ and $B(y) = B^{(1)}(y) = y$. For these functions $A(x) = x$ and $B(y) = y$, the expression (11) becomes the usual expression (3) for the Pearson's correlation coefficient r. So, for the usual Gaussian case, the above algorithm converges and leads to the desired result.

Important mathematical subtleties.

1°. There are cases when the above algorithm – and even the definition (9) – do not lead to the desired result.

For example, if $Y = X$ when $X \geq 0$ and $Y = X - Z^2$ for $X < 0$, where Z is a random variable which is independent of X, then the maximum in (9) is attained when we take $A(x) = x$ for $x \geq 0$, $A(x) = 0$ for $x \leq 0$, and similarly, $B(y) = y$ for $y \geq 0$ and $B(y) = 0$.

For these functions $A(x)$ and $B(y)$, we have $A(X) = B(Y)$ and thus, $L(A, B) = 1$. This value seems to indicate that X and Y are perfectly correlated, but in reality, they are only correlated when $X \geq 0$ and $Y \geq 0$ and they are definitely not well correlated when $X < 0$ and $Y < 0$.

This counterintuitive feature of the definition (9) appeared because we allowed functions $A(x)$ and $B(y)$ which are constant on some intervals. To avoid this counterintuitive feature, it is therefore reasonable to make sure that functions $A(x)$ and $B(y)$ are never constant. The functions $A(x)$ and $B(y)$ are supposed to be non-decreasing. Non-decreasing means that the derivative is non-negative, while constant means derivative is 0. Thus, it makes sense to select a small positive number $\delta > 0$ and, in the definition (9), only consider functions for which $A'(x) \geq \delta$ and $B'(y) \geq \delta$ for all x and y.

2°. Another important issue is the existence of the functions $A(x)$ and $B(y)$ which maximize $L(A, B)$. In general, a continuous function is guaranteed to attain its maximum value on a given domain D only if this domain is *compact*. A known Ascoly-Arzela theorem states that a compact class of functions should be uniformly continuous; for smooth functions, this means that there should be an upper bound M on the derivatives, such that $A'(x) \leq M$ and $B'(y) \leq M$ for all x and y.

3°. Because of Comments 1 and 2, it makes sense to fix two positive real numbers $\delta < M$ and to restrict ourselves only to functions $A(x)$ and $B(y)$ for which $\delta \leq A'(x) \leq M$ and $\delta \leq B'(y) \leq M$.

5 Case of Heavy-Tailed Distribution

Need to go beyond Pearson's correlation coefficient. Pearson's correlation coefficient r, as defined by the formula (3), implicitly assumes that the margianl distributions for X and Y have finite variance. In reality, however, many econometric-related distributions are *heavy-tailed*, with infinite variance. Let us show how we can extend the above definitions to the heavy-tailed case. For that, we first need to briefly recall the need for heavy-tailed distributions.

Heavy-tailed distributions are ubiquitous. In many practical situations, e.g., in economics and finance, we encounter heavy-tailed probability distributions, i.e., distributions for which the variance is infinite. These distributions surfaced in the 1960s, when Benoit Mandelbrot, the author of fractal theory, empirically studied the fluctuations and showed [14] that larger-scale fluctuations follow the power-law distribution, with the probability density function $\rho(x) = A \cdot x^{-\alpha}$, for some constant $\alpha \approx 2.7$. For this distribution, variance is infinite.

The above empirical result, together with similar empirical discovery of heavy-tailed laws in other application areas, has led to the formulation of *fractal theory*; see, e.g., [15, 16].

Since then, similar heavy-tailed distributions have been empirically found in other financial situations [2–4, 7, 17, 19, 24, 26–28], and in many other application areas [1, 9, 15, 18, 23].

Utility: reminder. People's economic behavior is determined by their preferences. A standard way to describe preferences of a decision maker is to use the notion of *utility* u; see, e.g., [5, 6, 11, 13, 22]. According to decision theory, a user prefers an alternative for which the expected value $\sum\limits_{i=1}^{n} p_i \cdot u_i$ of the utility is the largest possible. Alternative, we can say that the expected value $\sum\limits_{i=1}^{n} p_i \cdot U_i$ of the *disutility* $U \stackrel{\text{def}}{=} -u$ is the smallest possible.

Disutility caused by probabilistic uncertainty. If we know the exact value of a quantity, then we can make an optimal decision based on this value. If we do not know the exact value – e.g., if we only know the probability distribution $\rho(x)$ on the set of all possible values – then we have to make a decision based on *some* value m. Since the actual value x is, in general, different from m, this decision is not as perfect as the decision based on the exact knowledge x.

For example, if we knew exactly what will be the future price x of a certain financial instrument (e.g., stock), then (after applying an appopriate future-related discount), we will be able to find the exact price that we are willing to pay for this instrument. In practice, we do not know this future price; at best, we know the probability of future value. As a result, we set up a price correponding to some "expected" value m.

- If the actual value x is smaller than our prediction m, then we overpay and thus, lose money on this transaction.

– If the actual value x is larger than m, this means that we may have missed an opportunity to invest in this instrument.

In both cases, the difference between the actual value x and the selected value m leads to disutility.

Let $U(d)$ denote the disutility cause by the difference $d = x - m$. When the value m has been selected, the average disutility is equal to $\int U(x-m)\cdot \rho(x)\,dx$. We select the value m for which this disutility is the smallest possible. The resulting minimal disutility is the disutility caused by the probabilistic uncertainty:

$$d_U(X) \stackrel{\text{def}}{=} \min_m E[U(X-m)] = \min_m \int U(x-m)\cdot \rho(x)\,dx. \tag{41}$$

What if x partly depends on a known quantity y? If the desired quantity x is somewhat dependent on another (known) quantity y, then, once we know y, we thus have more knowledge about x and hence, our uncertainty-caused disutility will decrease.

It is reasonable to take the percentage of this decrease as a measure of dependence between x and y.

Case of linear dependence. In this paper, we are interested in the case of linear dependence $x = a + b\cdot y$. A linear dependence is either increasing or decreasing.

If we expect the dependence to be increasing, then it makes sense to consider dependencies with $b \ge 0$. Among all such dependencies, we should select the values a and $b \ge 0$ for which the expected disuility $E[U(X-(a+b\cdot Y)]$ is the smallest possible. The resulting remaining disutility is equal to

$$d_U^+(X\,|\,Y) = \min_{a;b\ge 0} E[U(X-(a+b\cdot Y)] =$$

$$\min_{a;b\ge 0} \int U(x-(a+b\cdot y))\cdot \rho(x,y)\,dx\,dy. \tag{42}$$

The corresponding decrease $D_U^+(X\,|\,Y)$ in disutility can be thus estimated as

$$D_U^+(X\,|\,Y) \stackrel{\text{def}}{=} \frac{d_U(X) - d_U^+(X\,|\,Y)}{d(X)}. \tag{43}$$

Similarly, if we expect the dependence of x on y to be decreasing, we should consider dependencies with $b \le 0$. Among all such dependencies, we should also select the values a and $b \le 0$ for which the expected disuility $E[U(X-(a+b\cdot Y)]$ is the smallest possible. The resulting remaining disutility is equal to

$$d_U^-(X\,|\,Y) = \min_{a;b\le 0} E[U(X-(a+b\cdot Y)] =$$

$$\min_{a;b\le 0} \int U(x-(a+b\cdot y))\cdot \rho(x,y)\,dx\,dy. \tag{44}$$

The corresponding decrease $D_U^-(X\,|\,Y)$ in disutility can be thus estimated as

$$D_U^-(X\,|\,Y) \stackrel{\text{def}}{=} \frac{d_U(X) - d_U^-(X\,|\,Y)}{d_U(X)}. \tag{45}$$

How is this idea related to Pearson's correlation coefficient? For the case when $U(d) = d^2$, as one can easily check:

– the optimal value m is the mean of the random variable X: $m = E[X]$;
– the corresponding value $d_U(X)$ is equal to the variance $V(X)$;
– for $r \geq 0$, the decrease $D_U^+(X \mid Y)$ is equal to r^2; and
– for $r \leq 0$, the decrease $D_U^-(X \mid Y)$ is equal to r^2.

How to make the above definition depend only on a copula. The above-defined values $D_U^{\pm}(X \mid Y)$, in general, change when we apply non-linear re-scalings $A(x)$ and $B(y)$ to the quantities X and Y. Similarly to the case of Pearson's correlation coefficient, to define a scale-independent quantity, a quantity depending only on the copula, it makes sense to take a maximum over all possible re-scaling. Thus, we arrive at the following definition.

Resulting definition. Let a disutility function $U(d)$ be given. For a joint distribution of two random variables X and Y, the corresponding measures of linearity L_U^+ and L_U^- will be defined as

$$L^+ = \max_{A(x),B(y)} D_U^+(A(X) \mid B(Y)), \quad L^- = \max_{A(x),B(y)} D_U^-(A(X) \mid B(Y)), \tag{46}$$

where maximum is taken over all possible non-decreasing functions $A(x)$ an $B(y)$, and the values $D_U^{\pm}$ are defined by the formulas (41)–(45).

Comment. Similarly to the case of Pearson's correlation coefficient, we can make the definition explicitly depending only on the copula $C(x, y)$ if we use the above formulas with the probability density $\rho(x, y) = \dfrac{\partial^2 C(x, y)}{\partial x \partial y}$.

Acknowledgments. This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, by Grants 1 T36 GM078000-01 and 1R43TR000173-01 from the National Institutes of Health, and by a grant N62909-12-1-7039 from the Office of Naval Research.

References

1. Beirlant, J., Goegevuer, Y., Teugels, J., Segers, J.: Statistics of Extremes: Theory and Applications, Wiley, Chichester (2004)
2. Chakrabarti, B.K., Chakraborti, A., Chatterjee, A.: Econophysics and Sociophysics: Trends and Perspectives, Wiley-VCH, Berlin (2006)
3. Chatterjee, A., Yarlagadda, S., Chakrabarti, B.K.: Econophysics of Wealth Distributions, Springer-Verlag Italia, Milan (2005)
4. Farmer, J.D., Lux, T. (eds.): Applications of statistical physics in economics and finance, a special issue of the Journal of Economic Dynamics and Control, 32(1), 1–320 (2008)

5. Fishburn, P.C.: Utility Theory for Decision Making, John Wiley & Sons Inc., New York (1969)
6. Fishburn, P.C.: Nonlinear Preference and Utility Theory, The John Hopkins Press, Baltimore, (1988)
7. Gabaix, X., Parameswaran, G., Vasiliki, P., Stanley, H.E.: Understanding the cubic and half-cubic laws of financial fluctuations, Physica A, 324, 1–5 (2003)
8. Gelfand, I.M., Fomin, S.V.: Calculus of Variations, Dover, New York (2000)
9. Gomez, C.P., Shmoys, D.B.: Approximations and Randomization to Boost CSP Techniques, Annals of Operations Research, 130, 117–141 (2004)
10. Jaworski, P., Durante, F., Härdle, W.K., Ruchlik, T. (eds.), Copula Theory and Its Applications, Springer Verlag, Berlin, Heidelberg, New York (2010)
11. Keeney, R.L., Raiffa, H.: Decisions with Multiple Objectives, John Wiley and Sons, New York (1976)
12. Kreinovich, V., Lakeyev, A., Rohn, J., Kahl, P.: Computational Complexity and Feasibility of Data Processing and Interval Computations, Kluwer, Dordrecht (1997)
13. Luce, R.D., Raiffa, R.: Games and Decisions: Introduction and Critical Survey, Dover, New York (1989)
14. Mandelbrot, B.: The variation of certain speculative prices, J. Business, 36, 394–419 (1963)
15. Mandelbrot, B.: The Fractal Geometry of Nature, Freeman, San Francisco, California (1983)
16. Mandelbrot, B., Hudson, R.L.: The (Mis)behavior of Markets: A Fractal View of Financial Turbulence, Basic Books (2006)
17. Mantegna, R.N., Stanley, H.E.: An Introduction to Econophysics: Correlations and Complexity in Finance, Cambridge University Press, Cambridge, Massachusetts (1999)
18. Markovich, N., ed.: Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice, Wiley, Chichester (2007)
19. McCauley, J.: Dynamics of Markets, Econophysics and Finance, Cambridge University Press, Cambridge, Massachusetts (2004)
20. McNeil, A.J., Frey, R., Embrechts, P.: Quantitative Risk Management: Concepts, Techniques, Tools. Princeton University Press, Princeton, New Jersey (2005)
21. Nelsen, R.B.: An Introduction to Copulas. Springer Verlag, Berlin, Heidelberg, New York (1999)
22. Raiffa, H.: Decision Analysis, Addison-Wesley, Reading, Massachusetts (1970)
23. Resnick, S.I.: Heavy-Tail Phenomena: Probabilistic and Statistical Modeling, Springer-Varlag, New York (2007)
24. Roehner, B.: Patterns of Speculation - A Study in Observational Econophysics, Cambridge University Press, Cambridge, Massachusetts, (2002)
25. Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures, Chapman and Hall/CRC, Boca Raton, Florida (2011)
26. Stanley, H.E., Amaral, L.A.N., Gopikrishnan, P., Plerou, V.: Scale invariance and universality of economic fluctuations, Physica A, 283, 31–41 (2000)
27. Stoyanov, S.V., Racheva-Iotova, B., Rachev, S.T., Fabozzi, F.J.: Stochastic models for risk estimation in volatile markets: a survey, Annals of Operations Research, 176, 293–309 (2010)
28. Vasiliki, P., Stanley, H.E.: Stock return distributions: tests of scaling and universality from three distinct stock markets, Physical Review E: Statistical, Nonlinear, and Soft Matter Physics, 77(3), Pt. 2, Publ. 037101 (2008)