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 Introduction: Traditional Approach to Detecting Linear Dependence

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., [10].

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., [21]). For a 2-D distribution with a cumulative distribution function $F(x, y) = \text{Prob}\,(X \le x \,\&\, Y \le y)$ corresponding to probability density $\rho(x, y)$, Pearson's correlation coefficient is defined as

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

where

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

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

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

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

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

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

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

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

In the case of an exact linear dependence $Y = c_0 + c_1 \cdot X$, this coefficient $r(F)$ is equal to 1 if $c_1 > 0$ and to -1 if $c_1 < 0$. Vice versa, if $r(F) = \pm 1$, this means that with probability 1, we have $Y = c_0 + c_1 \cdot X$ for appropriate coefficients c_0 and c_1.

In general, values $r(F) \neq 0$ indicate that there is an approximate linear dependence – and the closer $|r(F)|$ to 1, the closer is the the actual dependence to a linear one.

Validating a linear model. The square $R^2 = (r(F))^2$ is used, in statistics, as a "measure of fit" which is used to validate the linear model: the closer this square to 1, the better the fit.

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

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

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., [9, 17, 18].

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

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

Main idea behind the new definition. We consider a situation in which we know the copula $C(a, b)$ but we do not know the marginal distributions $F_X(x)$ and $F_Y(y)$. We would like to know whether there exist some marginal distributions for which the dependence between the corresponding random variables x and y is linear, i.e., for which, for which, for the corresponding probability distribution $F(x, y) = C(F_X(x), F_Y(y))$, the Pearson's coefficient is equal either to 1 or to -1.

For different marginal distributions, we have different values of the Pearson's correlation coefficient. The possibility to have $r(F) = 1$ for *at least one* pair of the marginal distributions means that the *maximum* L^+ of $r(F)$ over all pairs of possible marginal distributions is equal to 1. Thus, we can use this maximum to gauge to what extent a given copula represents an increasing linear dependence.

Similarly, the possibility to have $r(F) = -1$ for *at least one* pair of marginal distributions means that the *minimum* L^- of $r(F)$ over all such pairs is equal to -1. Thus, we can use this minimum to gauge to what extent a given copula represents a decreasing linear dependence. So, we arrive at the following definition.

Definition. *Let a copula $C(a, b)$ be given. By* measures of linearity *corresponding to this copula, we mean the values*

$$L^- \stackrel{\text{def}}{=} \min_{F_X(x), F_Y(y)} r(C(F_X(x), F_Y(y))); \tag{9a}$$

$$L^+ \stackrel{\text{def}}{=} \max_{F_X(x), F_Y(y)} r(C(F_X(x), F_Y(y))), \tag{9b}$$

where $r(F)$ denote Pearson's correlation coefficient (3) *corresponding to $F(x,y) = C(F_X(x), F_Y(y))$, and the minimum and maximum are taken over all possible marginal probability distributions $F_X(x)$ and $F_Y(y)$.*

Thus defined values L^- and L^+ depend only on the copula. In the above definition, we fix a copula $C(a,b)$, and we consider all possible 2-D probability distributions $F(x,y)$ corresponding to this copula. Therefore, the above-defined values L^- and L^+ depend only on the copula.

The values L^- and L^+ describe the possibility of a linear dependence. If $L^+ = 1$, this means that there exist marginal distributions $F_X(x)$ and $F_Y(y)$ for which $r(F) = 1$, i.e., for which the corresponding random variables X and Y are linearly related by an increasing linear dependence $Y = c_0 + c_1 \cdot X$, with $c_1 > 0$. Similarly, if $L^- = -1$, this means that the exist marginal distributions $F_X(x)$ and $F_Y(y)$ for which $r(F) = -1$, i.e., for which the corresponding random variables X and Y are linearly related by a decreasing linear dependence $Y = c_0 + c_1 \cdot X$, with $c_1 < 0$.

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 the approximate dependence to a linear one.

How to define the corresponding measure of fit. For validating a linear model, as a measure of fit M, it is reasonable to take the largest possible value of the traditional measure of fit $R^2 = (r(F))^2$ over all possible probability distributions corresponding to the given copula.

If the largest value of $(r(F))^2$ is attained when $r(F) > 0$, then $L^+ \geq |L^-|$, and the above-defined measure of fit is equal to $(L^+)^2$. If the largest value of $(r(F))^2$ is attained when $r(F) < 0$, then $|L^-| \geq L^+$, and the above-defined measure of fit is equal to $(L^-)^2$. These two cases can be combined into a single formula

$$M = \max((L^-)^2, (L^+)^2).$$

How to actually compute L^- and L^+ based on $F(x,y)$: an idea. A direct application of the above definition based on the known probability distribution $F(x,y)$ seems computationally expensive: first, we need to compute the copula, and then, based on this copula, we need to solve two optimization problems. It turns out that it is possible to compute L^- and L^+ more efficiently.

This possibility is related to the fact that, once we know a joint distribution $F(x,y)$ for non-discrete random variables X and Y (i.e., for random variables for which the corresponding marginal distributions $F_X(x)$ and $F_Y(y)$ are continuous functions), we can explicitly describe all other random variables (X', Y') with the same copula as (X, Y).

Indeed, by definition of the copula, for the original random pair (X, Y), we have $F(x,y) = C(F_X(x), F_Y(y))$. Thus, we have $C(a,b) = F(F_X^{-1}(a), F^{-1}(b))$, where $F^{-1}(x)$ denotes an inverse function. Since the pair (X', Y') is described by the same copula $C(a,b)$ as the pair (X, Y), the distribution function $F'(x', y')$ for this pair has the form $F'(x', y') = C(F_{X'}(x'), F_{Y'}(y'))$, where

$F_{X'}(x')$ and $F_{Y'}(y')$ are the corresponding marginal distributions. Substituting the above expression for the copula $C(a,b)$ into this formula, we conclude that $F'(x',y') = F(a(x'),b(y'))$, where we denoted $a(x') \stackrel{\text{def}}{=} F_X^{-1}(F_{X'}(x'))$ and $b'(x') \stackrel{\text{def}}{=} F_X^{-1}(F_{X'}(x'))$.

By definition of a cumulative distribution function $F(x,y) =$ $\text{Prob}\,(X \le x \,\&\, Y \le y)$, the formula $F'(x',y') = F(a(x'),b(y'))$ means that $\text{Prob}\,(X' \le x' \,\&\, Y' \le y') = \text{Prob}\,(X \le a(x') \,\&\, Y \le b(y'))$.

Since the cumulative distribution functions are non-decreasing, the inverses $F_X^{-1}(a)$ and $F^{-1}(b)$ are also non-decreasing and thus, the compositions $a(x')$ and $b(y')$ are also non-decreasing. So, the condition $X \le a(x')$ is equivalent to $A(X) \le x'$, where $A(x)$ denotes an inverse function to $a(x)$, and similarly the condition $Y \le a(y')$ is equivalent to $B(Y) \le y'$, where $B(y)$ denotes an inverse function to $b(y)$. Thus, we conclude that $\text{Prob}\,(X' \le x' \,\&\, Y' \le y') =$ $\text{Prob}\,(A(X) \le x' \,\&\, B(Y) \le y')$. In other words, the probability distribution of the pair (X',Y') is exactly the same as the probability distribution of the pair $(A(X),B(Y))$.

Vice versa, one can easily check that if we take any two strictly increasing functions $A(x)$ and $B(y)$, then for the pair (X',Y') with $X' = A(X)$ and $Y' =$ $B(Y)$, we get the exact same copula as for the original pair (X,Y).

In other words, all possible probability distributions (X',Y') corresponding to the same copula $C(a,b)$ as the pair of random variables (X,Y) can be obtained by considering appropriate non-decreasing transformations $X' = A(X)$ and $Y' = B(Y)$. For the variables, mean, variance, covariance, and correlation can be explicitly determined in terms of the functions $A(x)$ and $B(y)$. Thus, we arrive at the following easier-to-compute equivalent formulas for describing the desired measures of linearity L^- and L^+.

Towards an easier-to-compute equivalent definition of L^- and L^+. Let (X,Y) be random variables corresponding to a copula $C(a,b)$. Then, the measures of linearity L^- and L^+ can be computed as

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

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 $r(F)$, we conclude that

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

where

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

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

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

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

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

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

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

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

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

Comment. Strictly speaking, the above equivalence between copulas and non-linear re-scalings requires that we consider only strictly increasing functions $a(x)$ and $b(y)$, for which the inverses $A(x)$ and $B(y)$ are also strictly increasing. However, one can easily show that any non-decreasing function $A(x)$ can be approximated, with any given accuracy, by a strictly increasing one: e.g., we can approximate $A(x)$ by $A(x) + \varepsilon \cdot x$ for a sufficiently small $\varepsilon > 0$. Thus, in (10), it does not matter whether we take only strictly increasing functions or all non-decreasing ones.

Explicit expressions for L^- and L^+ in terms of the copula. The above equivalent reformulation was intended for the case when we still need to compute the copula. However, even when we already know the copula $C(a,b)$, the above reformulation can still simplify computations.

Indeed, the formula (9c) can be applied to any probability distribution corresponding to a given copula. In particular, it is well known that the copula itself is a probability distribution on the box $[0,1] \times [0,1]$, corresponding to uniform marginal distributions $F_X(x) = \text{Prob}\,(X \le x) = x$ and $F_Y(y) = \text{Prob}\,(Y \le y) = y$. For this probability distribution, $F(x,y) = C(x,y)$ and thus, $\rho(x,y) = \dfrac{\partial^2 C(x,y)}{\partial x \partial y}$. For this probability density, we can apply the above formulas (10)–(15), and compute the desired values L^- and L^+.

4 How to Actually Compute L^- and L^+

Analysis of the problem. In accordance with the above idea, for computing L^- and L^+, we will use the easier-to-compute equivalent reformulation (10) of the original definition of these two measures of linearity.

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., [7].

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

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

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

Subtracting 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 functions, we use the formula (11) to compute the desired value L^+.

Comment. As a result of this algorithm, we get functions A and B which minimize and maximize the expression (9c), and we have already shown that the resulting minimum L^- and maximum L^+ depend only on the copula. Thus, the *result* of applying this algorithm depends only on the copula – and do not depend on the marginal distributions.

However, since we start with *some* distribution $\rho(x, y)$ corresponding to the given copula, the conditional expectations computed on each iteration will be, in general, *different*. In other words, if we start with the distributions $F(x, y)$ corresponding to different marginal distributions $F_X(x)$ and $F_Y(y)$, then:

- on each iteration, we get *different* functions, but
- for all starting distributions (X, Y) corresponding to the same copula, in the limit (after all the iterations) we get functions $A(x)$ and $B(y)$ for which the distribution of the pair $(X', Y') = (A(X), B(Y))$ is *the same* – namely, the distribution which, among all distributions corresponding to the given copula, maximizes (or minimizes) the Pearson correlation coefficient $r(F)$.

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

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

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

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 $r(F)$. 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 Ascoli-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 $\delta \le A'(x) \le M$ and $\delta \le B'(y) \le M$.

5 Case of Heavy-Tailed Distribution: Second Related Problem

Need to go beyond Pearson's correlation coefficient. Pearson's correlation coefficient $r(F)$, as defined by the formula (3), implicitly assumes that the marginal 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 [12] that larger-scale fluctuations follow the power-law distribution, with the probability density function $\rho(y) = A \cdot y^{-\alpha}$, 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., [13, 14].

Since then, similar heavy-tailed distributions have been empirically found in other financial situations [2–4, 16, 22, 23], and in many other application areas [1, 8, 13, 15, 20].

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, 11, 19]. 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 $\rho(y)$ on the set of all possible values – then we have to make a decision based on *some* value m. Since the actual value y is, in general, different from m, this decision is not as perfect as the decision based on the exact knowledge y.

For example, if we knew exactly what will be the future price y of a certain financial instrument (e.g., stock), then (after applying an appropriate 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 corresponding to some "expected" value m.

– If the actual value y is smaller than our prediction m, then we overpay and thus, lose money on this transaction.
– If the actual value y 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 caused by the difference $d = y - m$. When the value m has been selected, the average disutility is equal to $\int U(y-m)\cdot \rho(y)\,dy$. 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:

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

What if y partly depends on a known quantity x? If the desired quantity y is somewhat dependent on another (known) quantity x, then, once we know x, we thus have more knowledge about y 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 $y = c_0 + c_1 \cdot x$. A linear dependence is either increasing or decreasing.

If we expect the dependence to be increasing, then it makes sense to consider dependencies with $c_1 \geq 0$. Among all such dependencies, we should select the values c_0 and $c_1 \geq 0$ for which the expected disuility $E[U(Y - (c_0 + c_1 \cdot X)]$ is the smallest possible. The resulting remaining disutility is equal to

$$d^+_U(Y \mid X) = \min_{c_0; c_1 \geq 0} E[U(Y - (c_0 + c_1 \cdot X)] =$$

$$\min_{c_0; c_1 \geq 0} \int U(y - (c_0 + c_1 \cdot x)) \cdot \rho(x,y)\,dx\,dy. \tag{42}$$

The corresponding decrease $D^+_U(Y \mid X)$ in disutility can be thus estimated as

$$D^+_U(Y \mid X) \stackrel{\text{def}}{=} \frac{d_U(Y) - d^+_U(Y \mid X)}{d(Y)}. \tag{43}$$

Similarly, if we expect the dependence of y on x to be decreasing, we should consider dependencies with $c_1 \leq 0$. Among all such dependencies, we should also select the values c_0 and $c_1 \leq 0$ for which the expected disuility $E[U(Y - (c_0 + c_1 \cdot X)]$ is the smallest possible. The resulting remaining disutility is equal to

$$d^-_U(Y \mid X) = \min_{c_0; c_1 \leq 0} E[U(Y - (c_0 + c_1 \cdot X)] =$$

$$\min_{c_0;c_1\le 0} \int U(y-(c_0+c_1\cdot x))\cdot \rho(x,y)\,dx\,dy. \tag{44}$$

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

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

How is this idea related to Pearson's correlation coefficient? It turns out that the Pearson's correlation coefficient $r(F)$ corresponds to the quadratic disutility function $U(d)=d^2$. Specifically, for the case when $U(d)=d^2$, as one can easily check:

- the optimal value m is the mean of the random variable Y: $m=E[Y]$;
- the corresponding value $d_U(Y)$ is equal to the variance $V(Y)$;
- for $r(F)\ge 0$, the decrease $D^+_U(Y\,|\,X)$ is equal to $R^2=(r(F))^2$; and
- for $r(F)\le 0$, the decrease $D^-_U(Y\,|\,X)$ is equal to $R^2=(r(F))^2$.

How to modify the above definition so that it depends only on the copula. Let us assume that we have a copula $C(a,b)$ and a disutility function $U(d)$. We can then define the corresponding measures of linearity L^- and L^+ as the maximum, correspondingly, of the expression $D^-_U(Y\,|\,X)$ or of the expression $D^+_U(Y\,|\,X)$ over all possible probability distributions $F(x,y)=C(F_X(x),F_Y(y))$ corresponding to the given copula $C(a,b)$.

This definition clearly depends only on the copula (and not on the marginal distributions).

An easier-to-compute equivalent reformulation. Similarly to the case of the Pearson's correlation coefficient, we can show that the above definitions can be reformulated in an easier-to-compute equivalent form. Namely, for a joint distribution of two random variables X and Y, the above measures of linearity L^-_U and L^+_U can be equivalently defined as

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

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

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. 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)
7. Gelfand, I.M., Fomin, S.V.: Calculus of Variations, Dover, New York (2000)
8. Gomez, C.P., Shmoys, D.B.: Approximations and Randomization to Boost CSP Techniques, Annals of Operations Research, 130, 117–141 (2004)
9. Jaworski, P., Durante, F., Härdle, W.K., Ruchlik, T. (eds.), Copula Theory and Its Applications, Springer Verlag, Berlin, Heidelberg, New York (2010)
10. Kreinovich, V., Lakeyev, A., Rohn, J., Kahl, P.: Computational Complexity and Feasibility of Data Processing and Interval Computations, Kluwer, Dordrecht (1997)
11. Luce, R.D., Raiffa, R.: Games and Decisions: Introduction and Critical Survey, Dover, New York (1989)
12. Mandelbrot, B.: The variation of certain speculative prices, J. Business, 36, 394–419 (1963)
13. Mandelbrot, B.: The Fractal Geometry of Nature, Freeman, San Francisco, California (1983)
14. Mandelbrot, B., Hudson, R.L.: The (Mis)behavior of Markets: A Fractal View of Financial Turbulence, Basic Books (2006)
15. Markovich, N., ed.: Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice, Wiley, Chichester (2007)
16. McCauley, J.: Dynamics of Markets, Econophysics and Finance, Cambridge University Press, Cambridge, Massachusetts (2004)
17. McNeil, A.J., Frey, R., Embrechts, P.: Quantitative Risk Management: Concepts, Techniques, Tools. Princeton University Press, Princeton, New Jersey (2005)
18. Nelsen, R.B.: An Introduction to Copulas. Springer Verlag, Berlin, Heidelberg, New York (1999)
19. Raiffa, H.: Decision Analysis, Addison-Wesley, Reading, Massachusetts (1970)
20. Resnick, S.I.: Heavy-Tail Phenomena: Probabilistic and Statistical Modeling, Springer-Varlag, New York (2007)
21. Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures, Chapman and Hall/CRC, Boca Raton, Florida (2011)
22. 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)
23. 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)