"Generalized Measures of Correlation for Asymmetry, Nonlinearity, and Beyond": Some Antecedents on Causality

David E. Allen[a,*], and Michael McAleer[b]

[a] *School of Mathematics and Statistics, University of Sydney, Department of Finance, Asia University, Taiwan, and School of Business and Law, Edith Cowan University, Australia*

[b] *Department of Finance, College of Management, Asia University, Taiwan, Discipline of Business Analytics, University of Sydney Business School, Australia, Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, The Netherlands, Department of Economic Analysis and ICAE, Complutense University of Madrid, Spain, Department of Mathematics and Statistics, University of Canterbury, New Zealand, and Institute of Advanced Sciences, Yokohama National University, Japan*

Abstract

This note comments on the Generalized Measure of Correlation (GMC) that was suggested by Zheng et al. (2012). The GMC concept was partly anticipated in some publications over 100 years earlier by Yule (1897, 1900, 1903, 1907, 1909) in the proceedings of the Royal Society, and by Kendall (1946). Other antecedents discussed include work on dependency by Renyi (1959) and Doksum and Samarov (1995), together with the Yule-Simpson paradox. The GMC metric partly extends the concept of Granger causality, so that we consider causality, graphical analysis and alternative measures of dependency provided by copulas.

Keywords: Skewed correlation, Bravais formula, Generalised Measure of

*Corresponding author
Email address: profallen2007@gmail.com (David E. Allen)

May 10, 2020

Correlation, Causality, Nonlinearity.
JEL Codes: C12, C100, C130.

1. Introduction

Zheng et al. (2012) suggest that Pearson's correlation, when used as a measure of explained variance, is well understood, but a major limitation is that it does not account for asymmetry. Zheng et al. (2012) present what they suggest is a broadly applicable correlation measure, and consider a pair of generalized measures of correlation (GMC) that deal with asymmetry in the explained variance, and linear or nonlinear relations between random variables.

The authors present examples under which the paired measures are identical, and become a symmetric correlation measure that is the same as the squared Pearson's correlation coefficient, so that Pearson's correlation is a special case of GMC. Zheng et al. (2012) suggest that the theoretical properties of GMC show that GMC can be applicable in numerous applications, and can lead to more meaningful conclusions and improved decision making.

Vinod (2015) applied the GMC metric in an economic paper which featured an analysis of development economic markets in a study of 198 countries, and also developed the R library package 'generalCorr' (2019). Allen and Hooper (2018) used the metric to analyse causal relations between the VIX, S&P500, and the realized volatility (RV) of the S&P500 sampled at 5-minute intervals. Chen et al. (2017) use a development of the GMC concept to suggest a new model-free feature screening approach, namely sure explained variability and independence screening (SEVIS). At the core of SEVIS is an application of GMC in the screening.

Zheng et al. (2012) intended to introduce effective and broadly applicable statistical tools for dealing with asymmetry and nonlinear correlations between random variables. For simplicity of illustration, they regard "linear" or "symmetric" as special cases of "nonlinear" or "asymmetric", respectively. In the case of "linear and symmetric," Pearson's correlation coefficient is an extremely important and widely-used analytical tool in statistical data analysis. Zheng et al. (2012, p.1537) claim that 'New dependence measures' that comprise Pearson's correlation coefficient as a special case should be of the greatest interest to practitioners.

The paper explores previous work related to this measure. There are a number of relevant themes in the prior work. Some of the related issues addressed by Zheng et al. (2012) were previously anticipated and developed in the *Proceedings of the Royal Society* by the British Statistician, Udney Yule in (1897), some 115 years earlier! This note sets out Yule's approach and gives Yule (1897, 1900, 1909) credit for covering some of the foundations of the topic, plus Yule (1903) and Pearson (1899), which anticipated the Yule-Simpson paradox.

Other relevant work includes that of Renyi (1959) on dependence, Kendall (1943, 1946) on regression, Kendall and Stuart (1979) on correlation ratios, and Doksum and Samarov (1995) on non-parametric estimation of global functions. In this comment, we consider some of these metrics, together with Granger causality, graphical analysis and the use of Vine copulas to capture dependencies.

2. Generalised Measure of Correlation

Zheng et al. (2012) point out that, despite its ubiquity, there are inherent limitations in the Pearson correlation coefficient when it is used as a measure of dependency. One limitation is that it does not account for asymmetry in the explained variances, which are often innate among nonlinearly dependent random variables. As a result, measures dealing with asymmetries are needed.

In order to meet this requirement, Zheng et al. (2012) developed Generalized Measures of Correlation (GMC). They commence with the familiar linear regression model, and the partitioning of the variance into explained and unexplained components:

$$Var\,(X) = Var(E(X \mid Y)) + E(Var(X \mid Y)), \tag{1}$$

whenever $E(Y^2) < \infty$ and $E(X^2) < \infty$. Note that $E(Var(X \mid Y))$ is the expected conditional variance of X given Y, so that $E(Var(X \mid Y))/Var(X)$ can be interpreted as the explained variance of X by Y. Thus, we can write:

$$\frac{E(Var(X \mid Y))}{Var(X)} = 1 - \frac{E(Var(X \mid Y))}{Var(X)} = 1 - \frac{E[\{X - E(X \mid Y)\}^2]}{Var(X)}.$$

The explained variance of Y given X can be defined similarly. This leads Zheng et al. (2012) to define a pair of generalized measures of correlation

(GMC) as:

$$\{GMC(Y \mid X), GMC(X \mid Y)\} = \left\{1 - \frac{E[\{Y - E(Y \mid X)\}^2]}{Var(Y)},\ 1 - \frac{E[\{X - E(X \mid Y)\}^2]}{Var(X)}\right\}. \tag{2}$$

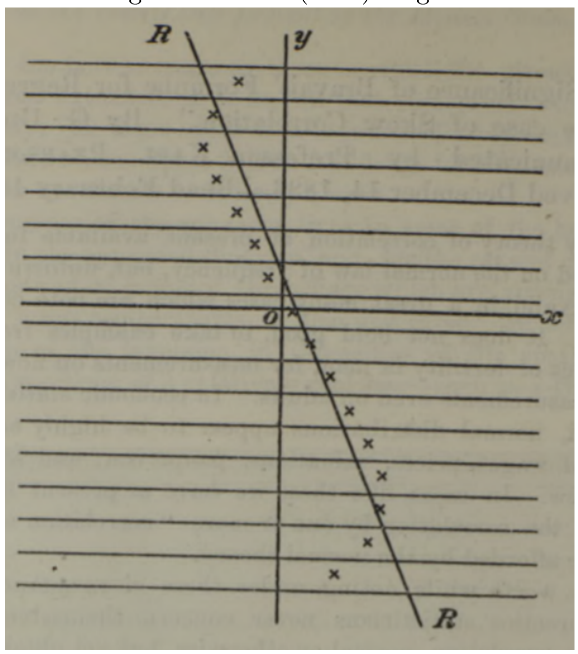
This pair of GMC measures has some attractive properties. It should be noted that the two measures are identical when (X, Y) is a bivariate normal random vector. However, GMCs are nonzero while Pearson's correlation coefficient may have a zero value when two random variables are nonlinearly dependent. GMC has various connections to Pearson's correlation coefficient and the coefficient of determination in regression models. They are identical to the squared Pearson's correlation coefficient when two random variables are related in a linear equation. A special case is where two random variables follow a bivariate normal distribution.

3. Yule's (1897) approach to general correlation

Yule (1897a, p.477) observed that: "The only theory of correlation at present available for practical use is based on the normal law of frequency, but, unfortunately, this law is not valid in a great many cases which are both common and important....in economic statistics, on the other hand, normal distributions appear to be highly exceptional: variation of wages, prices, valuations, pauperism, and so forth, are always skew."

He suggests letting Ox and Oy be the axes of a three-dimensional frequency surface drawn through the mean 0 of the surface parallel to the axes of

Figure 1: Yule's (1897) diagram



measurement, and the points marked (x) be the means of successive x arrays, lying on some curve that may be called the curve of regression of x on y. Then a line, RR, is fitted to this curve, as shown in Figure 1 (taken from his paper).

In commenting on the diagram, Yule notes that, if the slope of the line RR is positive, large values of x are associated with large values of y, while if negative, large values of x are associated with small values of y.

More importantly, for current purposes, Yule also notes that if the means of the arrays actually lie in a straight line (as in normal correlation), RR must be the equation to the line of the regression. Yule then lets n be the number of observations in any x array, and d be the horizontal distance of the mean of this array from the line RR. He then proposes to subject the line to the condition that the sum of all quantities like nd should be a minimum. In effect, he chooses to use the condition of least squares. He cautions that he does this solely for convenience of the analysis, and that he does not claim any advantages with regard to the probability of the results. He also cautions that it would be absurd to do so, as it is postulated at the outset that the curve of regression is only exceptionally a straight line, so that there can be no meaning in seeking the most probable straight line to represent the regression.

Yule proceeds by letting x and y be a pair of associated deviations, defines σ as the standard deviation of any array about its mean, and writes the equation of a straight line for RR as:

$$X = a + bY.$$

It follows that, for any one array:

$$S\{x - (a + by)\}^2 = S\{x - (a + bY\}^2 = n\sigma^2 + nd^2.$$

He extends the meaning of S to sum over the whole surface:

$$S(nd^2) = S\{x - (a + by)\}^2 - Sn\sigma^2,$$

where $Sn\sigma^2$ is independent of a, and is what he terms a characteristic of the surface. It follows that, if $S(nd^2)$ is set to a minimum, this is equivalent to making:

$$S\{x-(a+by)\}^2$$

a minimum. Yule suggests forming a single-valued relation:

$$x = a + by$$

between a pair of associated deviations, such that the sum of squares of errors in estimating any one x from the corresponding y is a minimum. This is simply the line of the regression RR. There will be two such equations to be formed corresponding to the two lines of the regression.

Yule then considers multiple combinations of variables that can be considered as two variables. As x and y represent deviations from their respective means, Yule suggests using S to denote summation over the whole surface:

$$S(x) = S(y) = 0.$$

The characteristic or regression equations are of the form:

$$\begin{aligned} x &= a_1 + b_1 y \\ y &= a_2 + b_2 x. \end{aligned} \tag{3}$$

Taking the equation for x, the normal equations for a_1 and b_1 are:

$$\begin{aligned} S(x) &= N(a_1) + b_1 S(y) \\ S(xy) &= a_1 S(y) + b_1 (Sy^2) \end{aligned} \tag{4}$$

with N being the number of correlated pairs. The first equation gives:

$$a_1 = 0,$$

while the second gives:

$$b_1 = \frac{S(xy)}{S(y^2)}.$$

Yule then lets $S(x^2) = N(\sigma_1^2)$, $S(y^2) = N(\sigma_2^2)$, and $S(x, y) = Nr\sigma_1\sigma_2$, where $\sigma_1$and σ_2 are the two standard deviations or mean square errors, and r is Bravais' (1846) value of the coefficient of correlation. Yule rewrites b_1 as:

$$b_1 = r\frac{\sigma_1}{\sigma_2}. \tag{5}$$

Similarly, when $a_2 = 0$:

$$b_2 = r\frac{\sigma_2}{\sigma_1}. \tag{6}$$

The expressions on the right of equations (3) and (4) are the values obtained by Bravais on the assumption of normal correlation for the regressions of x on y, and of y on x. Therefore, the Bravais values for the regressions are simply the values of b_1 and b_2 that make:

$$S(x - b_1y)^2 \quad and \quad S(x - b_2y)^2$$

their respective minima.

Denis (2000) observes that Bravais (1846) mathematically found the equation of the normal surface for the frequency of error. Using both analytic and geometric methods, Bravais also essentially found what would eventually be coined the "regression line", by investigating how the various elliptical areas of the frequency surface vary according to observed quantities. However, astronomers of the time were far more interested in "disposing" of this common error variance, largely due to the concern that errors would multiply, not compensate, when combining celestial observations.

Yule (1897) suggests proceeding by letting n be the number of correlated pairs in any one array taken parallel to the axis of x, and θ be the angle that the line of regression makes with the axis of y. For a single array:

$$S(xy) = yS(x) = ny^2 tan\theta,$$

or, extending S to summation over the whole surface:

$$S(xy) = N tan\theta \sigma_2^2,$$

or:

$$tan\theta = r\frac{\sigma_1}{\sigma_2}.$$

If the regression is linear, Bravais's formula may be used without investigating the normality of the distribution.

In the general case, both coefficients of regression must have the same sign, namely the sign of r. Hence, either regression will serve to indicate whether there is correlation or not. Yule suggests that the regressions are not convenient measures of correlation, as it may be found that:

$$b_1 > b_1^{'}, \quad b_2 < b_2^{'},$$

where b_1, b_2 and $b_1^{'} b_2^{'}$ are the regressions in the two cases. Yule queries to which distribution should we attribute the greater correlation? He observes that Bravais' coefficient solves the difficulty by taking the geometrical mean of the two regressions as the measure of correlation. It will still remain valid for non-normal correlations.

Yule generalizes the argument by suggesting that, instead of measuring x and y in arbitrary units, each is measured in terms of its own standard deviation:

$$\frac{x}{\sigma_1} = \rho \frac{y}{\sigma_2} \tag{7}$$

and solves for ρ by the method of least squares. A constant on the right-hand side can be ignored, as it would vanish, yielding:

$$\rho = \frac{S(xy)}{S(y^2)} \frac{\sigma_2}{\sigma_1} = r. \tag{8}$$

If measured x and y are each in terms of their respective standard deviations, r becomes the regression of x on y, and the regression of y on x.

Forming the sums of squares of the residuals in equations (1) and (6), and inserting the values of b_1, b_2, and ρ, gives:

$$S(x-b_1)^2 = N\sigma_1^2(1-r^2)$$
$$S(x-b_2)^2 = N\sigma_2^2(1-r^2) \quad (9)$$
$$S(\tfrac{x}{\sigma_1}-\rho\tfrac{y}{\sigma_2})^2 = S(\tfrac{y}{\sigma_2}-\rho\tfrac{x}{\sigma_1})^2 = N(1-r^2),$$

each of which is positive. Hence, r cannot be greater than unity. If r is equal to unity, each of the above becomes zero.

However,

$$S(\frac{x}{\sigma_1} \pm \frac{y}{\sigma_2})^2$$

can only vanish if:

$$\frac{x}{\sigma_1} \pm \frac{y}{\sigma_2} = 0$$

in every case, or if the following relation holds:

$$\frac{x_1}{y_1} = \frac{x_2}{y_2} = \frac{x_3}{y_3} = ... = \pm\frac{\sigma_1}{\sigma_2}, \quad (10)$$

with the sign of the last term in (10) dependent on the sign of r. Hence, the statement that two variables are perfectly correlated implies that relation (10) holds, or that all pairs of deviations bear the same ratio to one another. It follows that, where the means of the arrays are not collinear, or the deviation of the means of the arrays is not a linear function of the deviation, r cannot be unity. If the regression model is far from linear, caution must be used in using r to compare two different distributions. This caution is related to a central issue that Zheng et al. (2012) promote as one of the attractive properties of the GMC.

Equation (10) is the perfect case in the case of Gaussian distributions. If we take the ratios of random variables (observations) this may introduce nonlinear dependence. The explanation of GMC in equation [2] uses ratios (transformed) to construct a nonlinear dependence measure: the quotient correlation coefficient, as discussed in Zhang (2008)[1]. Equation (3) shows that the sample based Pearson's correlation coefficient and the quotient correlation coefficient are asymptotically independent and demonstrates its superior performance in testing hypothesis of independence. We return to a consideration of sample-based metrics when we consider Yule's work on partial correlations in section 5.

4. Kendall

Kendall (1946, p.145) devotes a section of his chapter of Volume II on regression estimation to a consideration of the fitting of curvilinear regression lines and comments that: "in general them means of arrays will not lie exactly on a smooth curve.... Nor do we know *a priori* what is the degree of a polynomial which will approximately represent the regression line". He then proceeds to assume that the regression line can be approximated by a polynomial of order p:

$$Y = a_0 + a_1 X + a_2 X^2 + ... + a_p X^p. \tag{11}$$

[1]We are grateful to a reviewer for drawing our attention to this point.

The problem is to determine the coefficients of a from the data. In effect, to find the values of the $a's$ that will minimise:

$$U = \sum (y - a_0 - a_1 x - .... - a_p x^p)^2 \tag{12}$$

with the sum extending over all sample values.

Differentiating with respect to a_j, he writes:

$$\sum (x^j y) - a_0 \sum x^j - a_1 \sum x^{j+1} - ... - a_p \sum x^{j+p} = 0, \tag{13}$$

with similar equations for $j = 0, ...p$. He then writes the moments without primes, for the sake of simplicity, and lets μ_j represent the jth moment of x, μ_{j1} the bivariate moment $\sum (x^j y)$, and writes:

$$\left.\begin{array}{c} a_0\mu_0 + a_1\mu_1 + ... + a_p\mu_p = \mu_{01} \\ a_0\mu_0 + a_1\mu_2 + ... + a_p\mu_{p+1} = \mu_{11} \\ ..... \\ a_0\mu_p + a_1\mu_{p+1} + ... + a_p\mu_{2p} = \mu_{p1} \end{array}\right\} \tag{14}$$

writing

$$\triangle^{(p)} = \begin{vmatrix} \mu_0 \ \mu_1 \ ....\mu_p \\ \mu_0 \ \mu_2 \ ....\mu_{p+1} \\ .... \\ \mu_0 \ \mu_{p+1} \ ....\mu_{2p} \end{vmatrix} \tag{15}$$

he then writes $\triangle_j^{(p)}$ for the determinant obtained by substituting the product moments $\mu_{01}....\mu_{p1}$ for the $j + 1$th column. He obtains, as the solution to

(13),

$$a_j = \frac{\triangle_j^{(p)}}{\triangle^{(p)}}. \tag{16}$$

Kendall (1946, p.146), notes that if the distribution function of the $x's$ is $G(x)$,we have for $\triangle^{(p)}$:

$$\triangle^{(p)} = \int\int ... \int \begin{vmatrix} 1 & x_0 & x_0^2 ... & x_0^p \\ x_1 & x_1^2 & x_1^3 ... & x_1^{p+1} \\ . & . & .... & . \\ x_p^p & x_p^{p+1} & x_p^{p+2} & x_p^{2p} \end{vmatrix} dG_0 dG_1 ... dG_p, \tag{17}$$

or if:

$$D = \begin{vmatrix} 1 & x_0 & ... & x_0^p \\ 1 & x_1 & ... & x_1^p \\ . & . & ... & ' \\ 1 & x_p & ... & x_p^p \end{vmatrix}. \tag{18}$$

$$\triangle^{(p)} = \int\int .... \int D^2 dG_0 dG_1 ... dG_p.$$

Kendall then permutes the suffixes of the $x's$ in all possible ways and sums the $(p+1)!$ resultants to obtain, by means of the definition of a determinant:

$$(p+1)! \triangle^p = \int\int ... \int D^2 dG_0 dG_1 ... dG_p ..., \tag{19}$$

and concludes that $\triangle^{(p)}$ is essentially positive.

From equations (17) and (13) it can be seen that the regression line can be written as:

$$\begin{vmatrix} Y & 1 & X... & X^p \\ \mu_{01} & \mu_0 & \mu_1... & \mu_p \\ \mu_{11} & \mu_1 & \mu_2... & \mu_p \\ . & . & . & . \\ \mu_{p1} & \mu_p & \mu_{p+1}... & \mu_{2p} \end{vmatrix} = 0. \quad (20)$$

This provides a formal solution to the problem and the moments μ can be obtained by observation, while equation (19) provides the regression line.

An alternative approach can yield the same solution. If it is assumed that the regression line is a parabolic curve of order p, then the coefficients may be obtained by the principle of moments. The lower moments could be obtained by:

$$\sum(x^j y) = \sum x^j(a_0 + a_1 x + ... + a_p x^p),$$

as far as was necessary to determine the $a's$, an approach which leads back to equation (15).

The only drawback is that we have no prior knowledge of the appropriate polynomial p required to fit the curve to the data. The only course of action is to fit curves of order $1, 2, 3, ...$ until we reach a point where adding further terms does not improve the fit. To avoid the problem of having to recalculate the determinant arithmetic every time we add a new term, we can consider then regression line in the following form:

$$Y = b_0P_0 + b_1P_1 + ... + b_pP_p, \tag{21}$$

where the Ps are polynomials in X, P_j being the degree of j. The Ps are determined so that:

$$\sum(P_jP_k) = 0, \;\; j \neq k, \tag{22}$$

with the summation extending over observed values.

The minimization of:

$$\sum(y - b_0P_0 - b_1P_1 - ...b_pP_p)^2, \tag{23}$$

produces equations such as:

$$\sum(yP_j) - b_0\sum(P_0P_j) - ... - b_p\sum(P_pP_j) = 0,$$

and, by virtue of the orthogonal relations in equation (21), this reduces to:

$$\sum(yP_j) - b_j\sum(P_j^2) = 0. \tag{24}$$

It follows that b_j is determined by P_j alone, and if a curve has been fitted of order p, and we wish to explore futher and add a further term $b_{p+1}P_{p+1}$,the coefficients $b_0, ..., b_p$, found from equation (23) remain unchanged.

Kendall (1946, p.147) further adds that the use of the orthogonal polynomials leads to a very convenient way of determining step by step, the goodness of fit of the regression.

We have:

$$U = \sum (y - b_0 P_0 - ... b_p P_p)^2$$

$$= \sum (y^2) - 2b_0 \sum (yP_0) - ... - 2b_p \sum (yP_p) + b_0^2 \sum (P_0^2) + ... + b_p^2 \sum (P_p^2).$$

Yet equation (23) shows that we can express $\sum (yP_j)$ in terms of $\sum (P_j^2)$, so that:

$$U = \sum (y^2) - b_0^2 \sum (P_0^2) - ... - b_P^2 \sum (P_P^2). \qquad (25)$$

This means that the effect of any term $b_j P_j$ is to reduce U by $b_j^2 \sum (P_j^2)$. It follows that if any additional term $b_p P_p$ does not significantly reduce U, it is redundant in terms of representing a regression line by a polynomial. This is exactly the approach typically taken in econometric software in the implementation of the Ramsey (1969) 'RESET' test.

Kendall (1946, p. 149) demonstrates in the explicit case of polynomials, (taking $\mu_1 = 0,\ \mu_2 = 1$), that:

$$P_0 = 1. \qquad (26)$$

$$P_1 = \frac{\begin{vmatrix} 1 & 0 \\ 1 & X \end{vmatrix}}{1} = X. \tag{27}$$

$$p_2 = \frac{\begin{vmatrix} 1 & 0 & 1 \\ 0 & 1 & \mu_3 \\ 1 & X & X^2 \end{vmatrix}}{\begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}} = X^2 - \mu_3 X - 1 \tag{28}$$

$$P_3 = \frac{\begin{vmatrix} 1 & 0 & 1 & \mu_3 \\ 0 & 1 & \mu_3 & \mu_4 \\ 1 & \mu_3 & \mu_4 & \mu_5 \\ 1 & X & X^2 & X^3 \end{vmatrix}}{\begin{vmatrix} 1 & 0 & 1 \\ 0 & 1 & \mu_3 \\ 1 & \mu_3 & \mu_4 \end{vmatrix}}$$

$$= \frac{1}{\mu_4 - \mu_3^2 - 1} \{(\mu_4 - \mu_3^2 - 1)X^3 - (\mu_5 - \mu_4\mu_3 - \mu_3)X^2$$

$$+(\mu_3\mu_5 - \mu_4^2 + \mu_4 - \mu_3^2)X + (\mu_5 - 2\mu_4\mu_3 + \mu_3)\}, \tag{29}$$

and so on. In the case where the population is normal:

$$P_1 = X$$

$$P_2 = X^2 - 1$$

$$P_3 = X^3 - 3X,$$

the polynomials reduce to the Tchebycheff-Hermite functions which produce an orthogonal system in the normal case.

In his work on partial correlation, Kendall (1943, p.385, notes and references), acknowledges the importance of Yule: "the theory of partial correlation is mainly due to Yule (1909)". Yule's formulation of the theory of partial correlation is considered in the next section.

5. Yule's formal analysis of partial correlation

The following discussion draws heavily on Aldrich (1998, p. 61), who comments on contributions by Gauss (1809, 1811), remarking that the process survives as Gaussian elimination without the associated notation, and that it did not enter the Pearson-Fisher mainstream of 20th Century statistics. He also notes the importance of a second scheme, namely correlation, that was introduced nearly a century later Yule (1909, p. 722), who saw it as 'an application to the purposes of statistical investigation' of least squares.

Aldrich (1998) suggests that Yule's notation was designed to improve on Karl Pearson's correlation notation that was used by Yule in his early work on the theory of partial and multiple correlation. For inference theory, Yule drew on Pearson (1896), and Pearson and Filon (1898). Its relation to Bayesian analysis is discussed in Aldrich (1997). Aldrich (1995) suggests that Yule's original system (1897, pp. 832-3) and presentation of partial correlation is based on a pair of lines fitted by least squares, written using deviations from the mean:

$$x_1 = b_{12}x_2 + b_{13}x_3$$
$$x_2 = b_{21}x_1 + b_{23}x_3$$

Yule defined ρ_{12}, as being the net, or partial correlation, between x_1 and x_2 with x_3 held constant, as:

$$\rho_{12} = \sqrt{b_{12}b_{21},} \qquad (30)$$

in an analogy of the relationship between the total correlation and the total regressions. The use of the above notation does not indicate which variable is being held constant.

Aldrich (1998, p.67) points out that Yule defined the partial correlations in terms of the slopes derived from least squares estimation, but used the normal equations rather than procedures from the least squares literature:

$$S(x_1x_2) = b_{12}S(x_2^2) + b_{13}S(x_2x_3)$$

$$S(x_1x_3) = b_{13}S(x_2x_3) + b_{13}S(x_3^2).$$

Yule then re-wrote the equations in terms of sample correlation coefficients $\rho_{12}, ...,$ etc, and sample standard deviations $\sigma_1, ...,$ and so on:

$$r_{12}\sigma_1 = b_{12}\sigma_2 + b_{13}r_{23}\sigma_3$$

$$r_{13}\sigma_1 = b_{12}r_{23}\sigma_2 + b_{13}\sigma_3.$$

These could be solved for b_{12}, for example, as:

$$b_{12} = \frac{r_{12} - r_{13}r_{23}}{(1 - r_{23}^2)} \cdot \frac{\sigma_1}{\sigma_2}. \tag{31}$$

Yule then used equations (21) and (20) and the corresponding formula for b_{21} to derive the following equation:

$$\rho_{12} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 = r_{13}^2)(1 - r_{23}^2)}}. \tag{32}$$

He was aware that the expression could become quite complex if there were a number of variables included as explanators in a regression equation.

Yule subsequently (1907, p. 182) developed his notation further, noting that: "the systems of notation hitherto used by writers on the theory of correlation are somewhat unsatisfactory when many variables are involved. In the present paper a new notation is proposed which is simple, definite, and quite general, thus very greatly facilitating the treatment of the subject".

In equation (31) ρ_{12}, is replaced by $r_{12.3}$, giving:

$$r_{12.34...n} = \frac{r_{12.4...n} - r_{13.4..n}r_{23.4..n}}{\sqrt{(1 - r_{13.4..n}^2)(1 - r_{23.4..n}^2)}}, \tag{33}$$

where $r_{12.34...n}$ is the partial correlation of x_1 and x_2 given $x_3, ...x_n$, and so on.

Yule (1907, p.184) suggests that: “this result is of some importance as regards the interpretation of partial correlations and regressions. In the case of normal correlation there is no difficulty in assigning a meaning to these constants, as the regression is strictly linear, and the partial correlations and regressions are the same for all types of the variables. But in the general case this is not so..”

This caution concerning the interpretation of partial correlation informed Yule’s analysis of causation. Aldrich (1995, p. 369) suggests that Yule summarised his view in his 1911 work introducing the theory of statistics. Yule opined that when an analysis of an association leads to the presumption of a direct causal relation when it does not exist, it is termed ’misleading’. This analysis was later taken up by Simon (1954, p. 467) in his investigation of ’spurious’ correlation.

Yule’s approach to partial correlation can be viewed as a sample based approach. Zhang (2008) proposed a quotient correlation which he defined as an alternative to Pearson’s correlation and as being more intuitive and flexible in cases where the tail behavior of data is important. He suggests that it measures nonlinear dependence where the regular correlation coefficient is generally not applicable and is a sample-based alternative to the Pearson correlation. Zhang at al. (2017) further explore the theoretical properties of the tail correlation coefficient (TQCC), which they proposed for measuring tail dependence between two random variables.

6. Renyi Dependence: Definitions and Notations

Renyi (1959) developed some fundamental properties capable of classifying measures of dependence. He suggests that $[\Omega, \eth, P]$ be a probability space, that is, Ω an arbitrary non-empty set whose elements will be denoted by ω, $\eth$, a σ-algebra of subsets of Ω whose elements will be denoted by capital letters $A, B, C, ..$, and $P = P(A)$, a probability measure on $\eth$. He denotes random variables on $[\Omega, \eth, P]$, that is real functions defined on Ω and measurable with respect to $\eth$, by Greek letters $\xi, \eta, ...$ If ξ is a random variable, we denote by $M(\xi)$ its mean value and by $D^2(\xi)$ its variance. If $M(\xi)$ and $D(\xi)$ exist and $D(\xi) > 0$, he defines:

$$\xi^* = \frac{\xi - M(\xi)}{D(\xi)}, \tag{34}$$

and calls the transformation by which ξ^* is obtained from ξ the standardization of ξ. If ξ is an arbitrary random variable, let $\eth_\xi$ denote the least σ-algebra of subsets of Ω with respect to which ξ is measurable. If η is another random variable with finite mean value, he denotes by $M(\eta \mid \xi)$ the conditional mean value of η with respect to a given value of ξ. $M(\eta \mid \xi)$ is a random variable, measurable with respect to $\eth_\xi$ and, is such that, for any $A \in \eth_\xi$, we have:

$$\int_A M(\eta \mid \xi) dP = \int_A \eta dP. \tag{35}$$

Of course, $M(\eta \mid \xi)$, is unique only if we consider two random variables which are equal with probability 1 to be identical. Renyi (1959) takes this for granted in his subsequent analysis. He then uses two well-known properties of conditional mean values:

$$M(M(\eta \mid \xi)) = M(\eta) \tag{36}$$

and:

$$M(g(\xi)\eta \mid \xi) = g(\xi)M(\eta \mid \xi), \tag{37}$$

if $g(x)$ is a Borel-measurable real function of the real variable x. The curve $y = M(\eta \mid \xi = x)$ is called the regression curve of η on ξ.

We shall denote the joint distribution of two random variables ξ and η by $Q_{\xi,\eta}$, that is, we put for any Borel subset C of the (x, y)-plane:

$$Q_{\xi,\eta}(C) = P((\xi, \eta) \in C),$$

where $Q_{\xi,\eta}(C)$ denotes the set of those $\omega \in \Omega$ for which the point with the coordinates $\xi(\omega), \eta(\omega)$ belongs to C. We denote by $Q_{\xi*\eta}$ the direct product of the distributions of ξ and η, that is, we put for any two Borel subsets A and B of the real line:

$$Q_{\xi*\eta}(A * B) = P(\xi \in A)P(\eta \in B),$$

where $A * B$ denotes the direct product of the sets A and B, that is, the set of all points (x, y) for which $x \in A$ and $y \in B$. The definition of $Q_{\xi*\eta}$ is extended to any Borel subset C of the (x, y)-plane.

6.1. Fundamental Properties of Measures of Dependence

Renyi (1959) then proceeds to consider the fundamental properties of measures of dependence and suggests we let ξ and η be random variables on a probability space $[\Omega, ð, P]$, neither of them being constant with probability 1. A common problem in the application of statistics is the need to characterize by a numerical value the strength of dependence between ξ and η. He suggests it is natural to choose a range between [0, 1] and to make the value 1 correspond to strict dependence and 0 to independence. Using the notation above, Renyi (1959), establishes the following postulates in relation to measures of dependence $\delta(\varepsilon, \eta)$:

- A) $\delta(\varepsilon, \eta)$ is defined for any pair of random variables ε and η, neither of them being constant with probability 1.
- B) $\delta(\varepsilon, \eta) = \delta(\eta, \varepsilon)$.
- C) $0 \leq \delta(\varepsilon, \eta) \leq 1$.
- D) $\delta(\varepsilon, \eta) = 0$, if ε and η are independent.
- E) $\delta(\varepsilon, \eta) = 1$ if there is strict dependence between ε and η, that is, either $\varepsilon = g(\eta)$ or $\eta = f(\varepsilon)$, where $g(x)$ and $f(x)$ are Borel-measurable functions.
- F) If the Borel-measurable functions $f(x)$ and $g(x)$ map the real axis in a one-to-one way onto itself, $\delta(f(\varepsilon)g(\eta)) = \delta(\varepsilon, \eta)$.

- G) If the joint distribution of ε and η is normal, then $\delta(\varepsilon,\eta) = | R(\varepsilon,\eta) |$, where $R(\varepsilon,\eta)$ is the correlation coefficient of ε and η.

Renyi (1959) proceeds to consider some individual measures of dependence in terms of these properties:

1. Correlation Coefficient. The correlation coefficient $R(\varepsilon,\eta)$ is defined, provided that $D(\varepsilon)$ and $D(\eta)$ are finite and positive, by:

$$R(\epsilon,\eta) = \frac{M(\varepsilon\eta) - M(\varepsilon)M(\eta)}{D(\varepsilon)D(\eta)} = M(\varepsilon^*,\eta^*). \qquad (38)$$

It has the range [-1,+1], so only its absolute value meets property C. Its absolute value also conforms to properties D and G, but not to the others.

7. Doksum and Samarov's Non-parametric Estimation of Global Functionals

Doksum and Samarov (1995) adopt a nonparametric regression setting and assess the asymptotic distributions of estimators of global integral functionals of the regression surface. They apply their results to the problem of obtaining reliable estimators for the nonparametric coefficient of determination. This coefficient, which is also called Pearson's correlation ratio, gives the fraction of the total variability of a response that can be explained by a given set of covariates. They show it can be used to construct measures of nonlinearity of regression and relative importance of subsets of regressors, and to assess the validity of other model restrictions.

Doksum and Samarov (1995) suggested that, for regression experiments where the relationship between a random covariate vector X and a response variable Y does not necessarily follow either a linear or other specified parametric model, a natural measure of the strength of the relationship between X and Y is Pearson's correlation ratio:

$$\eta^2 = \frac{Var(m(X))}{Var(Y)}, \tag{39}$$

where $m(x) = E(Y \mid X = x)$, $X \in R^d$, $Y \in R^1$. The correlation ratio η^2 is based on the ANOVA decomposition:

$$Var(Y) = Var(m(X)) + E(\sigma^2(X)), \tag{40}$$

where $\sigma^2(x) = Var(Y \mid X = x)$, and thus gives the fraction of the variability of Y which is explained with the best predictor based on X, $m(X)$. This can be interpreted as a nonparametric coefficient of determination or nonparametric R-squared. It can also be defined via the extremal correlation property:

$$\eta^2 = Corr^2(m(X), Y) = \sup_g Corr^2(g(X), Y), \tag{41}$$

where the supremum is taken over all real-valued functions $g(X)$ with finite second moments. Equation (39) can be proved using the iterated expectation property and the Cauchy-Schwarz inequality.

The quantity η^2 can also be interpreted in terms of signal-to-noise ratio, which is usually defined as the variability (or energy) of the signal X over that of the noise, $e = Y - m(X)$:

$$\frac{signal}{noise} = \frac{\eta^2}{1-\eta^2}.$$

Doksum and Samarov (1995) temporarily restrict attention to the case $d = dim(X) = 1$, and note that $\eta^2 = \eta^2_{xy}$, is an asymmetric measure. They suggest that it is possible that $\eta^2_{xy} = 1$, while $\eta^2_{yx} < 1$. Asymmetry of η^2 reflects the fact that it is a regression rather than a correlation measure of association. As such, it avoids some of the "pathologies" of ACE and the maximum correlation coefficient [see Renyi (1959), Breiman and Friedman (1985) and Buja (1990)].

The quantity η^2 is not a "strong" measure of association: while independence of X and Y clearly implies $\eta^2 = 0$; even $max(\eta^2_{xy}, \eta^2_{yx}) = 0$ does not imply that X and Y are independent. In fact, it is possible that $max(\eta^2_{xy}, \eta^2_{yx}) = 0$ while X and Y are functionally dependent, and they consider a uniform distribution on the unit circumference. They emphasize that X and Y may be dependent, though not through the conditional means, but in many other ways. On the other hand, if $Y = m(X) + \varepsilon$ with X and ε independent, $\eta^2 = 0$ is equivalent to the independence of X and Y.

They suggest that, in the nonparametric setup, estimates of the correlation ratio η^2 are quite sensitive to values of X near the boundary of its support S_X. By introducing a weight function, $\omega(X)$, which is equal to 1 in the central part of S_X, and is zero near the boundary of S_X, they obtain a more "robust" measure that focuses on the explanatory power of X without being too sensitive to values near the boundary. Doksum and Samarov (1995) consider the weighted functional η^2_ω, which they estimate using kernel regressions.

The above considerations are related to the issues addressed by Zheng et al. (2012) in their development of the GMC and their consideration of its appplication in tests of a form of Granger causality, as considered in section 9.

Other potential complications in both the study of dependency and the relations of cause and effect, are considered in the next section.

8. Yule-Simpson effect

The Yule-Simpson effect refers to a phenomenon in which a trend appears in several different groups of data, but disappears or reverses when these groups are combined. In discussing the challenges and opportunities provided by the emergence of 'big data', Nussbaum (2018, p. 489) mentions that: "the second concern is that at the heart of data analytics is the massive study of correlations. As we well know these correlations do not imply causation, so we can have misleading conclusions. We can also run into situations like Simpsons paradox and simply get the wrong answers."

Although Simpson (1951) re-drew attention to this issue, the concept had been covered earlier by Pearson et al. (1899) and Yule (1903). Pearson et al. (1899) considered the inheritance of fertility in man, and of fecundity in thoroughbred racehorses.

Yule (1903, p. 127), in the context of discussing inherited traits, states that: “the tests for independence are by no means simple when the number of attributes is more than two. Under what circumstances should we say that a series of attributes $ABCD$... were completely independent? I believe not a few statisticians would reply at once ’if the chance of finding them together were equal to the product of the chances of finding them separately,’ yet such a reply would be in error. The mere result:

$$\frac{(A, B, C, D, ...}{N} = \frac{(A)}{N} . \frac{(B)}{N} . \frac{(C)}{N} . \frac{(D)}{N}, \tag{42}$$

where N is the total number, does not in general give any information as to the independence or otherwise of the attributes concerned. If the attributes are known to be completely independent then certainly the relation (9) holds good, but the converse is not true”.

From the physical point of view, complete independence can only be said to subsist for a series of attributes $ABCD$... within a given universe when every pair of such attributes exhibits independence not only within the universe at large, but also in every sub-universe specified by one or more of the remaining attributes of the series, or their contraries.

This discussion leads to consideration of the role of causality in the context of time, as considered in section 10.

9. GMC in time series

Vinod (2015) applies the concept of GMC to analyse some concepts in development economics. He reminds the reader that GMC can be interpreted in terms of kernel causality, $GMC(Y \mid X)$, is the coefficient of determination, R^2, of the Nadaraya-Watson nonparametric kernel regression:

$$y = g(X) + \epsilon = E(Y \mid X) + \epsilon, \tag{43}$$

where $g(X)$ is a nonparametric, unspecified (nonlinear) function. Interchanging X and Y , we obtain the other $GMC(X \mid Y)$ defined as the R^2 of the Kernel regression:

$$X = g^{'}(Y) + \epsilon^{'} = E(X \mid Y) + \epsilon^{'}. \tag{44}$$

Vinod (2017) defines $\delta = GMC(X \mid Y) - GMC(X \mid Y)$ as the difference of two population R^2 values. When $\delta < 0$, we know that X better predicts Y than vice-versa. Hence, we define X kernel causes Y provided the true unknown $\delta < 0$. Its estimate $\delta^{'}$ can be readily computed by means of a regression.

Zheng et al. (2012) demonstrate that GMC can lead to a more refined version of the concept of Granger causality (see Granger (1969)). They assume an order one bivariate linear autoregressive model, wherein Y_t Granger causes X_t if:

$$E[\{X_t - E(X_t \mid X_{t-1})\}^2] > E[\{X_t - E(X_t \mid X_{t-1}, Y_{t-1})\}^2], \tag{45}$$

which suggests that X_t can be better predicted using the histories of both X_t and Y_t than using the history of X_t alone. Similarly, X_t Granger causes

Y_t if:

$$E[\{Y_t - E(Y_t \mid Y_{t-1})\}^2] > E[\{Y_t - E(Y_t \mid Yt-1, X_{t-1})\}^2]. \qquad (46)$$

They use the fact $E(Var(X_t \mid X_{t-1})) = E[\{X_t - E(X_t \mid X_{t-1}\}^2]$, and

$E[\{E(X_t \mid X_{t-1}) - E(X_t \mid X_{t-1}, Y_{t-1})\}^2] = E[\{X_t - E(X_t \mid X_{t-1})\}^2] - E[\{X_t - E(X_t \mid X_{t-1}, Y_{t-1})\}^2]$,

which suggests that equation (44) is equivalent to:

$$1 - \frac{E[\{X_t - E(X_t \mid X_{t-1}, Y_{t-1})\}^2]}{E(Var(X_t \mid X_{t-1}))} > 0. \qquad (47)$$

In the same way, (45) is equivalent to:

$$1 - \frac{E[\{Y_t - E(Y_t \mid Y_{t-1}, X_{t-1})\}^2]}{E(Var(Y_t \mid Y_{t-1}))} > 0. \qquad (48)$$

When both (44) and (45) are true, there is a feedback system.

Suppose that $\{X_{,t}, Y_t\}$, $Y_t > 0$ is a bivariate stationary time series. Zheng et al. (2012) define Granger causality generalized measures of correlation as:

$$GcGMC(X_t \mid \mathcal{F}_{t-1}) = 1 - \frac{E[\{X_t - \mid X_{t-1},\ X_{t-1}, ..., Y_{t-1}, Y_{t-2}, ...,)\}^2]}{E(Var(X_t \mid X_{t-1}, X_{t-2}, ...))}, \qquad (49)$$

$$GcGMC(Y_t \mid \mathcal{F}_{t-1}) = 1 - \frac{E[\{Y_t - \mid Y_{t-1},\ Y_{t-1}, ..., X_{t-1}, X_{t-2}, ...,)\}^2]}{E(Var(Y_t \mid Y_{t-1}, Y_{t-2}, ...))}, \qquad (50)$$

where $\mathcal{F}_{t-1} = \sigma(X_{t-1}, X_{t-2}, ..., Y_{t-1}, Y_{t-2}, ...)$.

Zheng et al. (2012) suggest the following:

- $GcGMC(X_t \mid \mathcal{F}_{t-1}) > 0$, Y Granger causes X.
- $GcGMC(Y_t \mid \mathcal{F}_{t-1}) > 0$, X Granger causes Y.
- $GcGMC(X_t \mid \mathcal{F}_{t-1}) > 0$ and $GcGMC(Y_t \mid \mathcal{F}_{t-1}) > 0$, there is a feedback system.
- $GcGMC(X_t \mid \mathcal{F}_{t-1}) > GcGMC(Y_t \mid \mathcal{F}_{t-1})$, X is more influential than Y
- $GcGMC(Y_t \mid \mathcal{F}_{t-1}) > GcGMC(X_t \mid \mathcal{F}_{t-1})$, Y is more influential than X.

10. GMC and causation

The potential use of GMC as a tool for exploring Granger causality leads naturally to consideration of the issues related to tests of causality. The classical philosopher, economist and historian, David Hume (1739, "SECT. XV. RULES BY WHICH TO JUDGE OF CAUSES AND EFFECTS"), had a great influence on the later development of the treatment of causality in economics. Hume discusses cause and effect in his 'Treatise of Human Nature', where he sets out eight propositions, four of which are given below:

"(1) The cause and effect must be contiguous in space and time.

(2) The cause must be prior to the effect.

(3) There must be a constant union betwixt the cause and effect. It is chiefly this quality, that constitutes the relation.

(4) The same cause always produces the same effect, and the same effect never arises but from the same cause. This principle we derive from experience, and is the source of most of our philosophical reasonings....".

Hume argues that our idea of necessary connection, which he concedes is the most characteristic element of causality, can arise only from our experience of the constant conjunction of particular temporal sequences. This suggests that causality has a very weak foundation.

Granger causality is an example of the modern probabilistic approach to causality, which is a natural successor to Hume. While Hume required constant conjunction of cause and effect, probabilistic approaches are content to identify cause with a factor that raises the probability of the effect: A causes B if $P(B \mid A) > P(B)$. The asymmetry of causality is secured by requiring the cause (A) to occur before the effect (B). Granger causality has been criticised because it is atheoretical.

A more inferential approach to the study of causality has recently been provided by the development of graph-theoretic approaches (see Pearl (2000, 2010)). This has been incorporated in dependence modelling via the application of copulas. Accounts of copula theory are available in Joe (1997) and Nelsen (2006). Hierarchical, copula-based structures have recently been used in some new developments in multivariate modelling. Notable among these structures is the pair-copula construction (PCC). Joe (1996) originally proposed PCC, and further exploration of its properties has been undertaken by Bedford and Cooke (2001, 2002) and Kurowicka and Cooke (2006). Aas et al. (2009) provided key inferential insights which have stimulated the use of the PCC in various applications, (see, for example, Chollete et al. (2009), Berg and Aas (2009), and Allen et al. (2013)).

Sklar (1959) provides the basic theorem describing the role of copulas for dependence in statistics, providing the link between multivariate distribution functions and their univariate margins. The argument proceeds as follows: let F be a d- dimensional distribution function with margins $F_1, ...F_d$. Then there exists a copula C such that, for all $x = (x_1, ..., x_d)' \in (\mathbb{R} \cup \{\infty, -\infty\})^d$,

$$F(x) = C(F_1(x_1), ..., F_d(x_d)). \tag{51}$$

C is unique if $F_1, ..., F_d$ are continuous. Conversely, if C is a copula and $F_1, ..., F_d$ are distribution functions, then the function F defined by (1) is a joint distribution with margins $F_1, ..., F_d$. In particular, C can be interpreted as the distribution function of a d-dimensional random variable on $[0, 1]^d$.

We can speak generally of the copula of continuous random variables $X = (X_1, ...X_d) \sim F$. The problem in practical applications is the identification of the appropriate copula.

Standard multivariate copulas, such as the multivariate Gaussian or Student-t, as well as exchangeable Archimedean copulas, lack the flexibility of accurately modelling the dependence among larger numbers of variables. Generalizations of these offer some improvement, but typically become rather intricate in their structure, and hence exhibit other limitations, such as parameter restrictions. Vine copulas do not suffer from any of these problems.

Initially proposed by Joe (1996), and developed in greater detail in Bedford and Cooke (2001, 2002), vines are a flexible graphical model for describing multivariate copulas built up using a cascade of bivariate copulas, so-called pair-copulas. Their statistical breakthrough was due to Aas, Czado, Frigessi, and Bakken (2009), who described statistical inference techniques for the two classes of canonical C-vines and D-vines. These belong to a general class of Regular Vines, or R-vines which can be depicted in a graphical theoretic model to determine which pairs are included in a pair-copula decomposition. Therefore, a vine is a graphical tool for labelling constraints in high-dimensional distributions.

A regular vine is a special case for which all constraints are two-dimensional or conditional two-dimensional. Regular vines generalize trees, and are themselves specializations of Cantor trees. Combined with copulas, regular vines have proven to be a flexible tool in high-dimensional dependence modelling. Copulas are multivariate distributions with uniform univariate margins. Representing a joint distribution as univariate margins plus copulas allows the separation of the problems of estimating univariate distributions from the problems of estimating dependence.

However, a drawback of a high dimensional vine copula is that after it has been constructed, one can integrate out the redundant variables to obtain a joint copula of the two variables of interest. Nevertheless, the resulting copula might be very different from the one constructed directly from a chosen copula family, in effect its uniqueness can not be guaranteed. Zheng et al. (2012, p. 1245) demonstrate that: “the bivariate GumbelHougaard copula is widely used in many applications, especially in finance and insurance. It is easy to show that the GMCs of a pair of random variables following a bivariate GumbelHougaard copula are identical”. This is not always the case though, as demonstrated by Zhang (2009), who showed that the three-sectional copula performed as well as the Gumbel-Hougaard copula in modelling bivariate extreme dependence, yet the three sectional copula was able to account for both symmetry and asymmetry in explained variances by varying coefficient values.

However, the focus in the development of the GMC is on the relationship between the means of two distributions, as captured by regression analysis. Graphical approaches and the application of copulas can be applied to capture dependencies across the entirety of two distributions. A more direct parallel between capturing dependencies via regression analysis would be quantile regression, as developed by Koenker and Bassett (1978), (for an extensive treatment see Koenker (2005)).

11. Conclusion

Zheng et al. (2012) provided a convincing explanation of the properties of the measure they refer to as a generalized measure of correlation (GMC). This comment draws attention to the fact that some of the properties of their suggested metric were explored previously by Yule (1897) in analysing skew correlation 115 years earlier, and in his subsequent work on partial correlation. Other antecedents presented for consideration include Renyi's (1959) consideration of dependence, and Doksum and Samarov's (1995) work on global functionals. This also raises the issue of the Yule-Simpson paradox and methods of confronting it.

The fact that the GMC metric also proposes a form of non-linear Granger causality led to considerations of causality, graphical analysis and copulas, as alternative methods of capturing dependencies, together with the causal analytical framework provided by Pearl (2000, 2010).

Acknowledgements

The authors are most grateful to the editors and three reviewers for very helpful comments and suggestions. The second author wishes to acknowledge the financial support of the Australian Research Council and the Ministry of Science and Technology (MOST), Taiwan.

References

[1] Aas, K., Czado, C., Frigessi, A., and H. Bakken (2009), Pair-copula constructions of multiple dependence, *Insurance, Mathematics and Economics*, 44:182–198.

[2] Aldrich, J. (1995), Correlations genuine and spurious in Pearson and Yule, *Statistical Science,* 10 (4): 364-376.

[3] Aldrich, J. (1997), R.A. Fisher and the making of Maximum Likelihood 1912-1922, *Statistical Science,* 12 (3): 162-176.

[4] Aldrich, J. (1998), Doing Least Squares: perspectives from Gauss and Yule, *International Statistical Review,* 66(1): 61-81.

[5] Allen, D.E and V. Hooper (2018), Generalized correlation measures of causality and forecasts of the VIX using non-linear models, *Sustainability*, 10(8: 2695), 1-15.

[6] Allen, D.E., A. Ashraf, M. McAleer, R.J. Powell, and A.K. Singh (2013), Financial dependence analysis: Applications of vine copulas, *Statistica Neerlandica*, 87, 4, 403-435.

[7] Bedford, T. and R. M. Cooke (2001), Probability density decomposition for conditionally dependent random variables modeled by vines, *Annals of Mathematics and Articial Intelligence*, 32, 245-268.

[8] Bedford, T. and R. M. Cooke (2002), Vines - a new graphical model for dependent random variables, *Annals of Statistics*, 30, 1031-1068.

[9] Berg, D. and K. Aas (2009), Models for construction of higher-dimensional dependence: A comparison study, *European Journal of Finance*, 15:639–659.

[10] Bravais, A. (1846), Analyse mathématique sur les probabilités des erreurs de situation d'un point, *Mémoires présentés par divers savants à l'Académie royale des sciences de l'Institut de France*, 9, 255-332.

[11] Breiman, L. and J. Friedman (1985), Estimating optimal transformations for multiple regression and correlation, *Journal of the American Statistical Association*, 80 580-598.

[12] Buja, A. (1990), Remarks on functional canonical variables, alternating least squares methods and ACE, *Annals of Statistics,* 18 1032-1069.

[13] Chen, M., Y.M. Lian, Z. Chen, and Z. Zhang, (2017), Sure explained variability and independence screening, *Journal of Nonparametric Statistics*, 29, 849-883.

[14] Chollete, L., A. Heinen, and A. Valdesogo (2009), Modeling international financial returns with a multivariate regime switching copula, *Journal of Financial Econometrics,* 7:437–480.

[15] Denis, D.J. (2000), The Origins of Correlation and Regression: Francis Galton or Auguste Bravais and the Error Theorists?, Paper presented at the 61st Annual Convention of the Canadian Psychological Association, Ottawa, Canada, 29 June, 2000.

[16] Doksum, K, and A. Samarov (1995), Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression, *Annals of Statistics*, 23(5), 1443-1473.

[17] Gauss, C.E (1809), *Theoria Motus Corporum Coelestium.* English translation by C.H. Davis, reprinted (1963), Dover, N York.

[18] Gauss, C.E (1811), *Disquisitio de Elementis Ellipticis Palladis.* English

translation of extract in pp. 148-155 of Trotter, H. EF (1957). Gauss's Work (1803-26) on the Theory of Least Squares, Technical Report 5, Statistical Techniques Research Group, Princeton University. A translation of Mithodes des Moindres CarrIs, the authorised French translation of Gauss's writings on least squares by J. Bertrand (1855), Paris: Mallet-Bachelier.

[19] Granger, C. (1969), Investigating causal relations by econometric methods and cross-spectral methods. *Econometrica,* 34, 424-438.

[20] Hume, D. (1739), *A Treatise of Human Nature,* reprinted Oxford, Clarendon Press, (1896).

[21] Joe, H. (1996), Families of m-variate distributions with given margins and m(m- 1)/2 bivariate dependence parameters, In L. Rüschendorf and B. Schweizer and M. D. Taylor, editor, *Distributions with Fixed Marginals and Related Topics.*

[22] Joe, H. (1997), *Multivariate Models and Dependence Concepts*, Chapman & Hall, London.

[23] Kendall, M.G. (1943), *The Advanced Theory of Statistics,* Charles Griffin, London.

[24] Kendall, M.G. (1946), *The Advanced Theory of Statistics,* Volume II, Charles Griffin, London.

[25] Kendall, M. and A. Stuart (1979), *The Advanced Theory of Statistics: Inference and Relationship*, Hodder Arnold, London.

[26] Koenker, R. and G. Bassett (1978), Regression Quantiles, *Econometrica*, 46(1) 33-50.

[27] Koenker, R. (2005), *Quantile Regression,* Econometric Society Monograph Series, Cambridge University Press.

[28] Kurowicka D. and R.M. Cooke (2003), A parametrization of positive definite matrices in terms of partial correlation vines, *Linear Algebra and its Applications*, 372: 225–251.

[29] Nelsen, R. (2006), *An Introduction to Copulas,* Springer, New York, 2nd edition

[30] Nussbaum, B.D. (2018), Statistics: essential now more than ever, *Journal of the American Statistical Association*, 113:522, 489-493.

[31] Pearl, J. (2000), *Causality: Models, Reasoning, and Inference,* Cambridge: Cambridge University Press.

[32] Pearl, J. (2010), The Foundations Of Causal Inference, *Sociological Methodology,* 40, 75–149.

[33] Pearson, K. (1896), Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmix, *Philosophical Transactions of the Royal Society A,* 187, 253-318.

[34] Pearson, K. and L.N.G. Filon, (1898), Mathematical Contributions to the Theory of Evolution IV. On the Probable Error frequency Constants and on the Influence of Random Selection on Variation and Correlation, Philosophical Transactions of the Royal Society A, 191, 229-311.

[35] Pearson, K, A. Lee, and L.Bramley-Moore (1899), Genetic (reproductive) selection: Inheritance of fertility in man, and of fecundity in thoroughbred racehorses, *Philosophical Transactions of the Royal Society Series A,* 192: 257–330.

[36] Ramsey, J.B. (1969) Tests for specification errors in classical Linear Least Squares Regression analysis, *Journal of the Royal Statistical Society. Series B*, 31(2): 350-371.

[37] A. Renyi (1959), On measures of dependence, *Acta Mathematica Academiae Scientiarum Hungarica*, 10(3-4) 441-451.

[38] Simon, H.A. (1954), Spurious correlation: A causal interpretation, *Journal of the American Statistical Association*, 49(267): 467-479.

[39] Simpson, E. H. (1951), The interpretation of interaction in contingency tables, *Journal of the Royal Statistical Society, Series B,* 13: 238–241

[40] Sklar, A. (1959), Fonctions de repartition a n dimensions et leurs marges, *Publications de l'Institut de Statistique de L'Universite de Paris*, 8, 229-231.

[41] Vinod, H.D. (2015), Generalized Correlation and Kernel Causality

with Applications in Development Economics, *Communications in Statistics - Simulation and Computation*, accepted Nov. 10, 2015, URL http://dx.doi.org/10. 1080/

[42] Vinod, H.D. (2019) R Package 'generalCorr', https://CRAN.R-project.org/package=generalCorr

[43] Yule, G.U. (1897a), On the Significance of Bravais Formulæ for Regression, in the case of skew correlation, *Proceedings of The Royal Society London*, 477-489.

[44] Yule, G.U. (1897b), On the Theory of Correlation, *Journal of the Royal Statistical Society*, 60, 812-854.

[45] Yule, G.U. (1900), On the association of attributes in statistics; with illustrations from the material of the childhood society, &c. *Philosphical Transactions Series A,* 194: 257-319.

[46] Yule, G.U. (1903), Notes on the theory of association of attributes in statistics, *Biometrika*, 2 (2): 121–134.

[47] Yule, G.U. (1907) On the Theory of Correlation for any Number of Variables, Treated by a New System of Notation, *Proceedings of the Royal Society of London, Series A*, 79, 182-193.

[48] Yule, G.U. (1909), The Applications of the Method of Correlation to Social and Economic Statistics. *Journal of the Royal Statistical Society*, 72, 721-730.

[49] Yule, G.U. (1911) *An Introduction to the Theory of Statistics,* Ist Edition, Griffin, London.

[50] Zhang, Z. (2008), Quotient correlation: a sample based alternative to Pearson's correlation, *Annals of Statistics*, 36, 1007-1030.

[51] Zhang, Z. (2009), On Approximating Max-Stable Processes and Constructing Extremal Copula Functions, *Statistical Inference for Stochastic Processes*, 12, 89–114.

[52] Zhang, Z., C. Zhang, and Q. Cui, (2017), Random threshold driven tail dependence measures with application to precipitation analysis, *Statistica Sinica*, 27, 685-709.

[53] Zheng, S., N-S, Shi and Z. Zhang (2012), Generalized measures of correlation for asymmetry, nonlinearity, and beyond, *Journal of the American Statistical Association*, 107, 1239-1252.