Hutchinson, Timothy Paul
Comment on "Trivariate statistical analysis of extreme rainfall events via the Plackett family of copulas" by Shih-Chieh Kao and Rao S. Govindaraju: How are trivariate copulas put to use? Water Resources Research, 2010; 46: W04801

Comment on "Trivariate statistical analysis of extreme rainfall events via the Plackett family of copulas" by Shih-Chieh Kao and Rao S. Govindaraju: How are trivariate copulas put to use?

T. P. Hutchinson[1]

Received 31 August 2009; revised 2 March 2010; accepted 8 March 2010; published 22 April 2010.

Citation: Hutchinson, T. P. (2010), Comment on "Trivariate statistical analysis of extreme rainfall events via the Plackett family of copulas" by Shih-Chieh Kao and Rao S. Govindaraju: How are trivariate copulas put to use?, *Water Resour. Res.*, *46*, W04801, doi:10.1029/2009WR008592.

1. Introduction

[1] It is suggested that the priorities for a formula for calculating a trivariate probability are that it is simple and easy to use, preferably expressed directly in terms of the univariate and bivariate marginal probabilities, and that it has a parameter that reflects trivariate association, in the sense that (starting from univariate and bivariate tail probabilities) different values of this lead to different trivariate probabilities. It is usually not a priority that the formula be a valid probability distribution. If this is accepted, simple formulae such as that proposed by *Kao and Govindaraju* [2008] may be of wider application than is evident, as the issue of compatibility of bivariate marginals can be put aside.

[2] When a (bivariate) distribution is transformed to uniform marginals, the result is called a copula. But when there are three variables, it is not clear how to construct a trivariate copula from the three bivariate copulas. The remarkable paper by *Kao and Govindaraju* [2008] makes a proposal. This comment first argues that their key idea may be of wider application than is evident (though this will depend on the purpose and setting of the use of a trivariate copula) and second discusses the interpretation of measures of trivariate association.

2. Priorities for a Trivariate Formula

[3] The task of finding how to construct a variety of different trivariate copulas from a given triplet of bivariate copulas is a difficult one. Given a triplet of formulae, it is usually not even known whether they are compatible. There is little understanding of what might be meant by trivariate association, and it is not known how much of it to expect or in what circumstances to expect more or less of it.

[4] *Kao and Govindaraju* [2008] have proposed assuming that a certain trivariate cross-product ratio, to which they give the symbol ψ, is constant (their equation (26)). They develop this idea in the context of the bivariate marginal distributions each having a constant cross-product ratio and extensively investigate the issue of compatibility of the bivariate marginals. Their proposed assumption is not, however, restricted to these bivariate marginals and could also be made for other triplets of bivariate copulas (as they recognize in paragraph 42). In that case the compatibility issue would reemerge.

[5] I wish to suggest, however, that for some applications (perhaps many), compatibility is a side issue and can be ignored. The priorities in calculating a trivariate probability are the following.

1. The formula is simple and easy to use, preferably expressed directly in terms of the univariate and bivariate marginal probabilities.

2. It has a parameter that reflects trivariate association. That is, given the univariate and bivariate tail probabilities, different values of this parameter lead to different trivariate probabilities.

[6] The first of the conditions makes it practicable to fit the formula to real trivariate data sets. It will then be possible to study what the trivariate association parameter is numerically, whether it tends to be bigger for some types of data than for others, and so on.

[7] Notice that I am not requiring that the formula be a valid probability distribution. Thus, if the formula is differentiated in order to get the probability density, the resulting expression may be negative in some regions. Expressed in another way, the trivariate tail probability might exceed one or more of the corresponding bivariate probabilities. (If that were found, one would substitute the bivariate probability for the trivariate.) Is that cheating? Perhaps, but the whole range of a distribution is typically not of equal interest. What is usually wanted is a trivariate tail probability, and a "probability density" that is negative in other parts of the distribution may not matter. Faced with intractable difficulties in proceeding conventionally, it seems reasonable to try this.

[8] The proposal by *Kao and Govindaraju* [2008] satisfies the second priority listed above and is a great advance on previous work (see their paragraphs 19–25). What would often be used is the trivariate normal distribution (after transformation of the empirical distributions to univariate normality), which is unsatisfactory as it is completely specified by the bivariate correlations and lacks any trivariate degree of freedom. As to the first priority listed above, solution of a quartic polynomial, equation (28) of Kao and Govindaraju, is not usually considered simple and straightforward, but it is perhaps simple and straightforward compared with other trivariate distributions.

[1]Centre for Automotive Safety Research, University of Adelaide, Adelaide, South Australia, Australia.

0043-1397/10/2009WR008592

[9] I have suggested a formula that might be a competitor to that of *Kao and Govindaraju* [2008]. (See *Hutchinson* [2010], whose work generalizes equation (8) of *Hutchinson* [1999], which applies to the special case of X and Y being independent.) Let S_{XYZ} be the trivariate survival function (the probability that X exceeds x, Y exceeds y, and Z exceeds z); let S_{YZ}, S_{XZ}, and S_{XY} be the three bivariate survival functions; and let S_X, S_Y, and S_Z be the three univariate survival functions. The formula is intended for use when the survival probabilities are quite small:

$$S_{XYZ}^{1/\alpha} = (S_X S_{YZ})^{1/\alpha} + (S_Y S_{XZ})^{1/\alpha} + (S_Z S_{XY})^{1/\alpha} - 2(S_X S_Y S_Z)^{1/\alpha}. \quad (1)$$

Here α may be regarded as a measure of trivariate tail behavior and might be approximately 7, on the basis of evidence from the trivariate normal distribution [*Hutchinson*, 1993]. This formula is not a valid trivariate distribution. It is, however, much simpler than the formula of *Kao and Govindaraju* [2008] (but that is partly offset by α being more difficult to estimate than ψ).

3. Setting and Purpose of the Calculation

[10] Using a formula that may not be a valid probability distribution, whether the formula is equation (1) above or that of *Kao and Govindaraju* [2008], raises the question of exactly what the setting and the purpose of the calculation are. A data set is summarized in a formula probably because there is insufficient data to use empirical probabilities directly. (Even if the data set is extensive, going back too far into the past means that things might have changed, and observations that are close together in time or space may not be independent.) It is also probably desired to make use of information about trivariate association, as otherwise, a procedure based on the trivariate normal distribution would be attractive. In general terms, then, a distribution or a formula is often intended to help in extrapolation: (1) extrapolation from the middle of the distribution (where there are many observations) to the tail of the distribution (where chief interest lies but where observations are sparse) and (2) extrapolation from what has been discovered in similar data sets from other places (about the shapes of distributions and the relationships between variables) to the data set being analyzed.

[11] Other relevant questions are what information is available as the starting point, what is to be calculated, what accuracy is required, how much time is available for the calculation, and (if the method is complicated) whether there is some way the result can be checked.

[12] It seems likely that a common intention is to estimate parameters in the formula and then to calculate probabilities in the tail of the distribution, where observations are too sparse to be reliable on their own. That is, at least one of X, Y, and Z is sufficiently extreme that, while there are enough data to use S_{YZ}, S_{XZ}, and S_{XY}, the empirically observed S_{XYZ} is unreliable. Instead, the trivariate parameter is first estimated, and then the calculation is made. Then, if several data sets are available, multiple comparisons of predicted and empirical S_{XYZ} can be made: even if the empirical values are individually unreliable because of few observations, the evidence in total about whether the predicted values tend to be too large or too small or about right may be considered useful. Instead of casting the comparison in terms of S_{XYZ}, ψ (or, alternatively, α) might be regarded as a descriptive property of a $2 \times 2 \times 2$ table of frequencies and might be calculated at several points in the region of the distribution that is of most interest (e.g., the upper tail). Suppose this were repeated for many data sets of the same type, such as rainfalls at three places. There would then be evidence about whether ψ (or α) really is constant within a given distribution, whether it tends to be related to abstract statistical properties of the distribution (e.g., at how extreme a point the triple dichotomization is made and how strongly associated the bivariate marginals are), whether it tends to be related to the physical variables (e.g., how much rainfall there is or the period over which rainfall has been aggregated), and so on.

4. Interpretation of Trivariate Association

[13] The commonsense idea of trivariate association is that all three variables tend to be large together or small together. However, contrary to this, when the bivariate marginals are fixed, a greater probability of the three variables all exceeding three thresholds necessarily implies a lower probability of their all being less than those thresholds. That is, suppose each variable of a trivariate distribution is dichotomized, and thus there is a $2 \times 2 \times 2$ table of frequencies. Let us try to increase trivariate association by increasing the frequency in the high deep right cell by an amount a. If we require that the bivariate marginals are unaltered, the three cells adjacent to the high deep right cell will change by an amount $-a$, the three cells adjacent to these will change by an amount a, and, finally, the low shallow left cell will change by an amount $-a$. That is, the frequency of the three variables all being less than their respective thresholds is reduced. *Kao and Govindaraju* [2008] say that their ψ does not have an intuitive interpretation. However, as ψ is based on dichotomizing each variable, it appears (in view of the foregoing argument) that ψ indicates where the trivariate association between the variables is located rather than how much there is. The same is true of α. *Kao and Govindaraju* [2008, Figure 2, left] also suggests this contrast between the upper tail and lower tail in that the distribution function K_C for $\psi = 20$ crosses that for $\psi = 1/20$.)

[14] It seems likely that a trivariate parameter that reflected the three variables being both large together and small together would need to be based on splitting each variable into three ranges, not two. That is, suppose each variable of a trivariate distribution is split into three ranges, and thus there is a $3 \times 3 \times 3$ table of frequencies. The cells might be labeled 000, 001, 002, 010, 011, …., 222. Now let us try to increase the frequency in cell 222 by an amount a and that in cell 122 by an amount b. The changes in the other cells can be worked out if it is required that the bivariate marginals are unaltered and also that the changes are symmetric both between the dimensions and between the upper tail and the lower tail. It is indeed possible for the frequency in cell 000 to also increase by an amount a; it turns out that the change to cell 111 is $4a + 6b$.

[15] This criticism of ψ and α suggests playing down the idea that equation (26) of *Kao and Govindaraju* [2008] represents a whole general purpose trivariate copula and instead emphasizing that ψ and α are tools, the usefulness of which should be judged in a specific setting such as the upper tail of the distribution (e.g., above the 70th percentiles of the variables), as described in section 3.

[16] **Acknowledgments.** The Centre for Automotive Safety Research receives core funding from both the South Australian Department for Transport, Energy and Infrastructure and the South Australian Motor Accident Commission. The views expressed are those of the author and do not necessarily represent those of the University of Adelaide or the funding organizations.

References

Hutchinson, T. P. (1993), The seventh-root formula for a trivariate normal probability, *Am. Stat.*, *47*(2), 102–103, doi:10.2307/2685186.

Hutchinson, T. P. (1999), Familial association of disease and the structure of trivariate distributions, *Ann. Hum. Genet.*, *63*(6), 539–544, doi:10.1046/j.1469-1809.1999.6360539.x.

Hutchinson, T. P. (2010), Discussion of "Fully nested 3-copula: Procedure and application on hydrological data" by F. Serinaldi and S. Grimaldi, *J. Hydrol. Eng.*, in press.

Kao, S.-C., and R. S. Govindaraju (2008), Trivariate statistical analysis of extreme rainfall events via the Plackett family of copulas, *Water Resour. Res.*, *44*, W02415, doi:10.1029/2007WR006261.

T. P. Hutchinson, Centre for Automotive Safety Research, University of Adelaide, Adelaide, SA 5005, Australia. (paul@casr.adelaide.edu.au)