Why Are FGM Copulas Successful: A Simple Explanation

Songsak Sriboonchitta[1] and Vladik Kreinovich[2]
[1]Faculty of Economics
Chiang Mai University
Chiang Mai, Thailand
songsakecon@gmail.com
[2]Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
vladik@utep.edu

Abstract

One of the most computationally convenient non-redundant ways to describe the dependence between two variables is by describing the corresponding copula. In many application, a special class of copulas – known as FGM copulas – turned out to be most successful in describing the dependence between quantities. The main result of this paper is that these copulas are the fastest-to-compute, and this explains their empirical success.

As an auxiliary result, we also show that a similar explanation can be given in terms of fuzzy logic.

Keywords: copula, FGM copula, computational complexity, fuzzy logic

1 Introduction

What is a copula: a brief reminder. In many practical situations, we know the distribution of each of the two random variables X and Y, and we now need to also describe their joint distribution.

The distribution of each of the random variables can be described by the corresponding cumulative distribution functions $F_X(x) \stackrel{\text{def}}{=} \text{Prob}(X \le x)$ and $F_Y(y) \stackrel{\text{def}}{=} \text{Prob}(Y \le y)$.

Similarly, to describe their joint distribution, we can use corresponding 2-D cumulative distribution function (cdf)

$$F_{XY}(x, y) \stackrel{\text{def}}{=} \text{Prob}(X \le x \,\&\, Y \le y).$$

In principle, we can thus try to determine the values $F_{XY}(x, y)$ corresponding to all possible pairs (x, y). However, from the practical viewpoint, this is redundant; indeed:

- the 2-D cdf $F_{XY}(x, y)$ also contains information about the 1-D cdfs $F_X(x)$ and $F_Y(y)$, as $F_X(x) = F_{XY}(x, +\infty)$ and $F_Y(y) = F_{XY}(+\infty, y)$,
- so if we determine all the values $F_{XY}(x, y)$, we will also be determining the values $F_X(x)$ and $F_Y(y)$, but
- we consider the cases when the 1-D cdf values are already known, so soliciting them again is unnecessary.

It is therefore desirable to describe the dependence between X and Y in a non-redundant way, so that:

- from this description, we will not be able to extract the known 1-D cdfs, but
- from this information and from the 1-D cfds, we will be able to extract the 2-D cdf.

Such a non-redundant description is indeed known, it is a *copula* $C(u, v)$, a function from $[0, 1] \times [0, 1]$ to $[0, 1]$ for which, for all real numbers x and y, we have

$$F_{XY}(x, y) = C(F_X(x), F_Y(y));$$

see, e.g., [4, 6, 8, 9, 12].

Properties of copulas. Not every function $C(u, v)$ is a copula for an appropriate 2-D distribution. For a function to be a copula, it has to satisfy some properties. In this paper, we will use the following properties – which can be easily derived from the definition of the copula:

$$C(0, v) = C(u, 0) = 0; \quad C(1, v) = v; \quad C(u, 1) = u. \tag{1}$$

FGM copulas and their success. There exist many different copulas. Interestingly, in many practical applications, the following Farlie-Gumbel-Morgenstern (FGM) copula turns out to be very successful

$$C(u, v) = u \cdot v + \theta \cdot u \cdot (1 - u) \cdot v \cdot (1 - v)$$

for $\theta \in [-1, 1]$. The original papers are [2, 3, 10]; see, e.g., [7, 14] and references therein for latest results.

Why? To the best of our knowledge, until now, there was no convincing explanation of why FGM copulas are so empirically successful. In this paper, we provide such an explanation.

2 Materials and Methods

2.1 Explanation Based on Computational Complexity: Main Result

Statistical data processing is computing. Statistical data processing involves a large amount of computing. With the ever increasing amount of data, processing all this data requires more and more computation time – often to the extent that we exceed the capabilities of our computers.

From this viewpoint, it is desirable to select techniques which are as computationally efficient as possible. With respect to copulas, this means that we should select copulas $C(u, v)$ whose values are the easiest (and thus, the fastest) to compute.

Which functions are the fastest to compute? In the computers, the only exactly hardware supported operations are addition, subtraction, and multiplication. Everything else – from division to special functions such as $\exp(x)$, $\sin(x)$, etc. – is approximated by a sequence of elementary hardware supported operations. The more accuracy we need, the more elementary operations we need, and thus, the longer the corresponding computations.

So, the fastest-to-compute functions are functions that can be exactly represented as a sequence of elementary operations: in this case, the number of elementary operations remains the same no matter what accuracy we desire in our computations. In other words, we are looking for functions which can be obtained from constants and original quantities $x_1, \ldots, x_n$ by applying addition, subtraction, and multiplication. One can easily see that such functions are *polynomials*; indeed:

- every polynomial is a sum of monomials, and each monomial is a product of a constant and variables, so each polynomial is indeed a superposition of additions and multiplications;
- vice versa, each constant and each variable are polynomials, and the sum, the difference, and the product of two polynomials is also a polynomial; thus, by induction, we can prove that every superposition of addition, subtraction, and multiplication is a polynomial.

Not all polynomials are equally easy or equally difficult to compute. Out of the three elementary operations, the most time-consuming operation is multiplication. Thus, the fewer multiplications, the faster is the computation of the corresponding function.

- With one multiplication – performed in parallel – we can compute linear functions $a_0 + \sum_{i=1}^{n} c_i \cdot x_i$, and also products $x_i \cdot x_j$ of two variables.
- By applying second multiplication to the results of the first one, we can thus compute 3rd degree polynomials – or products of 4 variables, etc.

In general, the higher the degree, the more time is needed to compute the corresponding polynomial.

Resulting idea. From the viewpoint of selecting fastest-to-compute copulas, we should select polynomial copulas, and among them – copulas of the smallest possible degree.

Let us describe the results of such a selection.

Proposition. *Every polynomial copula has the form*

$$C(u,v) = u \cdot v + \theta(u,v) \cdot u \cdot (1-u) \cdot v \cdot (1-v),$$

for some polynomial $\theta(u,v)$.

Comments.

- For reader's convenience, the proof is placed in the special proof section.
- As a consequence of this proposition, we get the following results.

Corollary 1. *The only polynomial copula of 3rd degree is* $C(u,v) = u \cdot v$.

Comment. This copula is actually of 2nd degree, it corresponds to the case of two independent variables. Thus, to describe dependence, we need to consider polynomials of higher degree.

Corollary 2. *The only polynomial copulas of 4th degree are FGM copulas.*

Comments.

- This result explains the empirical success of the FGM copulas: among copulas describing true dependence, they are the easiest to compute.
- Since the FGM copulas are symmetric $C(u,v) = C(v,u)$, asymmetric dependence requires higher-degree polynomial copulas.
- An alternative explanation of the FGM formulas, based on *fuzzy logic*, is given in the next subsection.

2.2 Explanation Based on Computational Complexity: Proof of the Main Result

1°. The first condition on the copula, the condition that $C(0,v) = 0$ for all v, means that if $u = 0$, then $C(u,v) = 0$.

An arbitrary polynomial $C(u,v)$ can be represented as

$$C(u,v) = C_0(v) + u \cdot C_1(u,v),$$

where $C_0(v)$ is the sum of all the monomials that do not contain u, and $C_1(u,v)$ is the result of dividing all u-containing monomials by u.

For $u = 0$, the condition $C(0,v) = 0$ means that $C_0(v) = 0$ for all v. Thus, $C(u,v) = u \cdot C_1(u,v)$ for some polynomial $C_1(u,v)$.

2°. The condition $C(u,0) = u \cdot C_1(u,0) = 0$ for all $u \neq 0$ implies that $C_1(u,0) = 0$ for all u, and thus, that $C_1(u,v) = v \cdot C_2(u,v)$ for some function $C_2(u,v)$. So,

$$C(u,v) = u \cdot C_1(u,v) = u \cdot v \cdot C_2(u,v).$$

3°. The condition $C(1,v) = v$ takes the form $v \cdot C_2(1,v) = v$, so $C_2(1,v) = 1$, and so $f(u,v) \stackrel{\text{def}}{=} C_2(u,v) - 1 = 0$ when $u = 1$, i.e., when $1 - u = 0$.

4°. Similarly to Part 1 of this proof, this implies that

$$C_2(u,v) - 1 = (1-u) \cdot C_3(u,v)$$

for some polynomial $C_3(u,v)$. Similarly, the condition $C(u,1) = 1$ implies that $C_3(u,v) = (1-v) \cdot C_4(u,v)$ for some polynomial $C_4(u,v)$. Thus,

$$C_2(u,v) - 1 = (1-u) \cdot C_3(u,v) = (1-u) \cdot (1-v) \cdot C_4(u,v),$$

hence

$$C_2(u,v) = 1 + (1-u) \cdot (1-v) \cdot C_4(u,v)$$

and

$$C(u,v) = u \cdot v \cdot C_2(u,v) = u \cdot v \cdot (1 + (1-u) \cdot (1-v) \cdot C_4(u,v)).$$

This is the desired formula, with $\theta(u,v) = C_4(u,v)$.

The proposition is proven.

2.3 Explanation Based on Fuzzy Logic

What is fuzzy logic: a brief reminder. An alternative explanation comes from *fuzzy logic*, where numbers from the interval $[0,1]$ describe the expert's degree of confidence in a statement. Fuzzy logic was invented by L. Zadeh [15]; for the state-of-the-art, see, e.g., [1, 5, 11, 13].

In fuzzy logic, once we know the expert's degree of confidence a in a statement A, his/her degree of confidence in its negation $\neg A$ is estimated as $1 - a$.

Similarly, if we know the expert's degree of confidence a in a statement A, and we know the expert's degree of confidence b in a statement B, then the expert's degree of confidence in a conjunction $A \wedge B$ is estimated as $f_\wedge(a,b)$ for an appropriate function $f_\wedge(a,b)$; this function is known as an *"and"-operation* or a *t-norm*. One of the most widely use "and"-operations is the algebraic product $f_\wedge(a,b) = a \cdot b$ – that corresponds to the situation when A and B are statistically independent and we take probability as degree of confidence. This is the "and"-operation that we will use in this appendix.

Similarly, to estimate the expert's degree of confidence in a statement $A \vee B$, we apply an appropriate "or"-operation $f_\vee(a,b)$ (also called *t-conorm*) to the corresponding degrees a and b. One of the most widely used "or"-operations

is $f_\vee(a,b) = \min(a+b,1)$. This is the "or"-operation that we will use in this section.

Copula as a particular case of an "and"-operation. A copula can also be viewed as an "and"-operation: it transforms the probabilities $F_X(x) = \text{Prob}(X \le x)$ and $F_Y(y) = \text{Prob}(Y \le y)$ of the events $X \le x$ and $Y \le y$ into the probability $F_{XY}(x,y) = \text{Prob}(X \le x \,\&\, Y \le y)$ that the first event occurs *and* the second event occurs. How can we go from the original "crisp" "and"-operation to a new "fuzzy" one?

Towards a fuzzy explanation of the FGM copula. For each of the two statements A and B, we want to cover both possibilities:

- that the corresponding statement is absolutely true, and
- that the corresponding statement is "fuzzy" – i.e., to some extent true and to some extent false.

In other words, fuzzy means that there is some degree of belief that A is true *and* that its negation is true.

Thus, we can say that the statement $A \,\&\, B$ is true if

- either A and B are absolutely true,
- or A and B are both "fuzzy" – i.e., true to some extent and false to some extent.

The degree to which A is true is a. Thus, the degree to which the negation $\neg A$ is true is $1-a$. So, the degree to which both the statement A and its negation are both true is $a \cdot (1-a)$. This is a degree to which the statement A is fuzzy.

Similarly, the degree to which B is fuzzy is equal to $b \cdot (1-b)$. Thus, the degree to which both A and B are fuzzy is equal to the product $a \cdot (1-a) \cdot b \cdot (1-b)$.

If we denote the degree to which this both-fuzzy case contributes to "and" by θ, then the contribution of this case to the overall true of the conjunction $A \,\&\, B$ is $\theta \cdot a \cdot (1-a) \cdot b \cdot (1-b)$.

The degree to which both A and B are true can be estimated as $a \cdot b$. So, if we use $\min(a+b,1)$ as the "or"-operation, then the resulting overall degree has the desired form

$$a \cdot b + \theta \cdot a \cdot (1-a) \cdot b \cdot (1-b);$$

(at least while this sum does not exceed 1 – and for the FGM copulas, it does not exceed 1).

So, we indeed have an alternative – fuzzy-logic-based – explanation of the FGM copula.

3 Discussion and Conclusion

Problem: reminder. In many practical applications, correlation is used to describe dependence between random variables. However, correlation only captures possible linear dependence between random variables. To describe a general – possibly nonlinear – dependence, we need to use, e.g., the copula techniques.

There exist many different families of copulas. It turns out that in many applications, the actual dependence between random variables is best described by copulas from a special family of FGM copulas. Up to now, there has been no convincing explanations for this empirical observation.

Our results. In this paper, we provide two possible theoretical explanations for this empirical phenomenon. First, we show that the FGM copulas are the easiest to compute – this is one possible explanation for their empirical success. Second, we show that these copulas naturally appear when we use fuzzy logic to formalize our imprecise understanding of how to describe the dependence between random variables.

Discussion. The fact that these two explanations lead to the same class of empirically successful copulas make us confident that this is indeed the best possible class.

Our results will also, hopefully, make practitioners and researchers more confidence that FGM copulas are indeed the best, and thus, encourage them to use these copulas even more.

Remaining open problems. An interesting open problem is related to the fact that the FGM family of copulas is a 1-parametric family. This family may be the most accurate approximator among all 1-parametric families, but the general dependence can be more complex that this. So, to get an even more accurate description of the dependence between several variables, it is desirable to use 2- and more-parametric families. Which 2-, 3-, ..., -parametric families should we use?

Can we use computational complexity-related ideas to come up with appropriate multi-dimensional families of copulas? Our arguments imply that all elements of such families should be polynomials of higher order, but what exactly formulas should we use? Can we use fuzzy logic to transform our informal understanding of this problem into precise formulas for such families? Or do we need new methods for that? This would be interesting to investigate. A good start would be to first analyze this problem empirically: which 2-parametric families of copula is empirically the best?

Acknowledgments

This work was supported by the Center of Excellence in Econometrics, Chiang Mai University, Thailand. It was also supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of

Excellence) and DUE-0926721, and by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation.

The authors are greatly thankful to all the participants of the 2017 International Conferences of the Thailand Econometric Society TES'2017, especially to Zheng Wei, for valuable discussions, and to the anonymous referees for important suggestions.

A preliminary version of this paper was posted at the University of Texas at El Paso Technical Report UTEP-CS-17-24.

Conflict of Interest Statement

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] R. Belohlavek, J. W. Dauben, and G. J. Klir, *Fuzzy Logic and Mathematics: A Historical Perspective*, Oxford University Press, New York, 2017.

[2] D. G. J. Farlie, "The performance of some correlation coefficients for a general Bivariate distribution", *Biometrika*, 1960, Vol. 47, pp. 307-323.

[3] E. J. Gumbel, "Bivariate exponential distributions", *Journal of the American Statistical Association*, 1960, Vol. 55, pp. 698-707.

[4] P. Jaworski, F. Durante, W. K. Härdle, and T. Rychlik (eds.), *Copula Theory and Its Applications*, Springer Verlag, Berlin, Heidelberg, New York, 2010.

[5] G. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic*, Prentice Hall, Upper Saddle River, New Jersey, 1995.

[6] V. Kreinovich, H. T. Nguyen, S. Sriboonchitta, and O. Kosheleva, "Why copulas have been successful in many practical applications: a theoretical explanation based on computational efficiency", In: V.-N. Huynh, M. Inuiguchi, and T. Denoeux (eds.), *Integrated Uncertainty in Knowledge Modeling and Decision Making, Proceedings of The Fourth International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making IUKM'2015*, Nha Trang, Vietnam, October 15-17, 2015, Springer Lecture Notes in Artificial Intelligence, 2015, Vol. 9376, pp. 112–125.

[7] V. Kreinovich, S. Sriboonchitta, and V. N. Huynh (eds.), *Robustness in Econometrics*, Springer Verlag, Cham, Switzerland, 2017.

[8] J.-F. Mai and M. Scherer, *Simulating Copulas: Stochastic Models, Sampling Algorithms, and Applications*, World Scientific, Singapore, 2017.

[9] A. J. McNeil, R. Frey, and P. Embrechts, *Quantitative Risk Management: Concepts, Techniques, and Tools*, Princeton University Press, Princeton, New Jersey, 2015.

[10] D. Morgenstern, "Einfache beispiele zweidimensionaler verteilungen", *Mitteilungsblatt für Mathematische Statistik*, 1956, Vol. 8, pp. 234–235.

[11] J. M. Mendel, *Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions*, Springer, Cham, Switzerland, 2017.

[12] R. B. Nelsen, *An Introduction to Copulas*, Springer Verlag, Berlin, Heidelberg, New York, 2007.

[13] H. T. Nguyen and E. A. Walker, *A First Course in Fuzzy Logic*, Chapman and Hall/CRC, Boca Raton, Florida, 2006.

[14] Z. Wei, D. Kim, T. Wang, and T. Tetranont, "A multivariate generalized FGM copula and its application to multiple regression", In: V. Kreinovich, S. Sriboonchitta, and V. N. Huynh (eds.), *Robustness in Econometrics*, Springer Verlag, Cham, Switzerland, 2017, pp. 363–380.

[15] L. A. Zadeh, "Fuzzy sets", *Information and Control*, 1965, Vol. 8, pp. 338-353.