FINDING CORRELATION OF FUZZY DATA

A.S. Wungreiphi, Fokrul A. Mazarbhuiya, and Limainla Kichu

[1]*Dept. of Mathematics, School of Fundamental and Applied Sciences, Assam Don Bosco University, Assam*

*For correspondence. (fokrul.mazarbhuiya@dbuniversity.ac.in)

Abstract: While dealing with the non-fuzzy data, finding statistical parameters like mean, variance, standard deviation and correlation coefficient between variables are very common. However, evaluating the value of such statistical parameters are not so straightforward in case of fuzzy data. In this article, we propose a new method to find the correlation coefficient of two fuzzy sets in terms of their membership values. For this purpose, we define covariance using the membership values of the fuzzy sets. We derive the formula for finding correlation coefficient for the fuzzy sets over both continuous and discrete universal sets. We have shown in this paper that the value of coefficient lies in the interval [0, 1]. We then demonstrate our work with the help of numerical examples

Keywords: Fuzzy numbers; Membership values; Covariance of fuzzy sets; Discrete sets; Set of real numbers; Mathematical induction

1. Introduction:

The body of the paper should appear like this. Figures and tables should be included here and not at the end. The table should be included as follows after giving a single spacing.

In the field of mathematics, Zadeh [1] developed the concept of fuzziness. As a result, a number of authors have looked into the mathematics of the fuzzy statistical parameters like, mean, median, standard deviation, correlation coefficient [2, 3, 4]. In [5], the authors have proposed a method to calculate the correlation coefficient for intuitionistic fuzzy by adopting the concept from the conventional statistics where an intuitionistic fuzzy set is defined in terms of membership and non-membership functions. The value of the correlation coefficient calculated by [5] not only provide the strength of the relationship of intuitionistic fuzzy sets but also shows that the intuitionistic fuzzy sets are positively or negatively related. In [6], the authors have defined the correlation of two intuitionistic fuzzy sets as sum of the products of membership values plus sum of the products of non-membership values. The definition given in [6] coincides with that of [7]. Extending their work in [8], the authors have not only introduced the concept correlation and correlation coefficients but also studied the properties of interval-valued intuitionistic fuzzy sets.

In [9], the authors have adopted a method from central interval to calculate the correlation coefficient for fuzzy data which most of the previous method did not do. The authors of [10] used a straightforward approach based on the usual concept of Pearson correlation coefficient to derive management strategy or choice using fuzzy interval measures. The authors of [11] have extended the traditional method to see if a set of multivariate fuzzy data can be explained by a multivariate normal distribution.

In most of aforesaid cases the universe of discourse is considered as finite discrete sets. However, finding correlation between fuzzy sets defined over continuous set ($\subseteq$R) of universe of discourse can be interesting. In this article we are proposing a single formula for both continuous and discrete cases. We have also shown that in both the cases the correlation coefficient value lies between 0 and 1 as in the case of most of the earlier work.

The paper is organized as follows. In Section-2, we discuss about problem definition. The proposed method is discussed Section-3 along with numerical examples. Finally, we conclude our paper with a brief conclusions and lines for future works in Section-4.

2. Problem Definition

Let us consider X to be the universe of discourse. A fuzzy set A is defined as follows
A={(x, A(x)); A(x) ∈ [0, 1], x∈X} [see e. g. [12]]
where A(x), the membership function representing the membership grade of x in A.

A fuzzy set A is called as normal if ∃ at least one x∈X, for which μA(x) =1. For a fuzzy set A, an α-cut Aα [12] is represented by Aα={x∈X; μA(x)≥α}. If all the α-cuts of A are convex sets then A is said to be convex.

A convex normal fuzzy set A on R (real line) with the property that ∃ an x0∈R such that μA(x0) =1, and μA(x), piecewise continuous is called fuzzy number. A special case of fuzzy number are triangular numbers.

Fuzzy intervals are types of fuzzy numbers such that ∃ [a, b]⊂R such that μA(x0)=1for all x0∈ [a, b], and μA(x) is piecewise continuous. A special case fuzzy intervals are trapezoidal numbers.

3. Proposed Method

The proposed of finding correlation coefficient of fuzzy sets defined over discrete or continuous universe of discourse is given as follows.
For the two fuzzy A ={(xi, A(xi)); xi∈X}, and B = {(xi, B(xi)); xi∈X}over a discrete universe of discourse X={x1, x2,…xn}, then we define the correlation coefficient between A and B by the given formula

$$\rho(A,B) = \frac{cov(A,B)}{\sqrt{cov(A,A).cov(B,B)}} \quad (1)$$

where $cov(A,B) = \sum_{i=0}^{n} A(x_i)B(x_i)$ = covariance of A and B, $cov(A,A) = \sum_{i=0}^{n}(A(x_i))^2$ = covariance of the same fuzzy set A, and $cov(B,B) = \sum_{i=0}^{n}(B(x_i))^2$= covariance of the same fuzzy set B.

For the two fuzzy A ={(x, A(x)); x∈X⊆R}, and B = {(x, B(x)); x∈X⊆R}over a continuous universe of discourse X, then we define the correlation coefficient between A and B by the given same formula

$$\rho(A,B) = \frac{cov(A,B)}{\sqrt{cov(A,A).cov(B,B)}} \quad (2)$$

whereas $cov(A,B) = \int_X A(x)B(x)dx$ = covariance of A and B, $cov(A,A) = \int_X A(x)^2dx$ = covariance of the same fuzzy set A, and $cov(B,B) = \int_X B(x)^2dx$= covariance of the same fuzzy set B.

Theorem 1. If A ={(x, A(x)); x∈X}, and B = {(x, B(x)); x∈X} are two fuzzy sets over a universe of discourse X (discrete or continuous), then the correlation coefficient $\rho(A,B)$ satisfies the following properties

i) If A=B, then ρ(A, B)=1

ii) ρ(A, B)= ρ(B, A)

iii) 0≤ρ(A, B)≤1.

Proof. Properties i) and ii) are evident from the definition (2).

To prove property iii) let us take discrete case. Let A={(xi, A(xi)); xi∈X}, and B = {(xi, B(xi)); xi∈X} are two fuzzy sets over a discrete universe of discourse X={x1, x2,…xn}. The correlation coefficient of A and B is given by (1) as follows

$$\rho(A,B) = \frac{cov(A,B)}{\sqrt{cov(A,A).cov(B,B)}}$$

$$= \frac{\sum_{i=0}^{n} A(x_i)B(x_i)}{\sqrt{\sum_{i=0}^{n}(A(x_i))^2.\sum_{i=0}^{n}(B(x_i))^2}} = \frac{A(x_1)B(x_1)+A(x_2)B(x_2)+\cdots+A(x_n)B(x_n)}{\sqrt{(A(x_1)^2+A(x_2)^2+\cdots+A(x_n)^2).(B(x_1)^2+B(x_2)^2+\cdots+B(x_n)^2)}} \quad (3)$$

Obviously $\rho(A,B)\geq 0$. It remains to show that $\rho(A,B) \leq 1$.

$\Rightarrow \frac{A(x_1)B(x_1)+A(x_2)B(x_2)+\cdots+A(x_n)B(x_n)}{\sqrt{(A(x_1)^2+A(x_2)^2+\cdots+A(x_n)^2).(B(x_1)^2+B(x_2)^2+\cdots+B(x_n)^2)}} \leq 1$ [using (3)]

$\Rightarrow (A(x_1)B(x_1)+A(x_2)B(x_2)+\cdots+A(x_n)B(x_n))2 \leq (A(x_1)^2+A(x_2)^2+\cdots+A(x_n)^2).(B(x_1)^2+B(x_2)^2+\cdots+B(x_n)^2)$ [using cross multiplication and then squaring]

$\Rightarrow (A(x_1)^2+A(x_2)^2+\cdots+A(x_n)^2).(B(x_1)^2+B(x_2)^2+\cdots+B(x_n)^2)-(A(x_1)B(x_1)+A(x_2)B(x_2)+\ \ldots+A(x_n)B(x_n))2 \geq 0$ (4)

If we can show that (4) is true then we can claim that $\rho(A,B) \leq 1$ is true. For this purpose, let us use Mathematical Induction.

For n=1, (4) is true and hence the result is obvious.

Putting n=2 in (4), we get

$$(A(x_1)^2+A(x_2)^2).(B(x_1)^2+B(x_2)^2)-\big(A(x_1)B(x_1)+A(x_2)B(x_2)\big)^2$$

$$=\big(A(x_1)B(x_2)-A(x_2)B(x_1)\big)^2 \geq 0$$

Thus the result (4) is true for n=2. Let us suppose (4) is true for n=k, therefore

$\Rightarrow (A(x_1)^2+A(x_2)^2+\cdots+A(x_k)^2).(B(x_1)^2+B(x_2)^2+\cdots+B(x_k)^2)-(A(x_1)B(x_1)+A(x_2)B(x_2)+\cdots+A(x_k)B(x_k))2 \geq 0$ (5)

We want show that (4) is true for n=k+1.

$(A(x_1)^2+A(x_2)^2+\cdots+A(x_k)^2+A(x_{k+1})^2).(B(x_1)^2+B(x_2)^2+\cdots+B(x_k)^2+B(x_{k+1})^2)-(A(x_1)B(x_1)+A(x_2)B(x_2)+\cdots+A(x_k)B(x_k)+A(x_{k+1})B(x_{k+1}))2$

$=[(A(x_1)^2+A(x_2)^2+\cdots+A(x_k)^2).(B(x_1)^2+B(x_2)^2+\cdots+B(x_k)^2)-(A(x_1)B(x_1)+A(x_2)B(x_2)+\cdots+A(x_k)B(x_k))2]+(A(x_1)^2+A(x_2)^2+\cdots+A(x_k)^2).(B(x_{k+1})^2)+(A(x_{k+1})^2).(B(x_1)^2+B(x_2)^2+\cdots+B(x_k)^2)+A(x_{k+1})^2B(x_{k+1})^2-2(A(x_1)B(x_1)+A(x_2)B(x_2)+\cdots+A(x_k)B(x_k)).A(x_{k+1})B(x_{k+1}))-A(x_{k+1})^2B(x_{k+1})^2$

$\geq (A(x_1)^2B(x_{k+1})^2-2A(x_1)B(x_{k+1})A(x_{k+1})B(x_1)+A(x_{k+1})^2B(x_1)^2+(A(x_2)^2B(x_{k+1})^2-2A(x_2)B(x_{k+1})A(x_{k+1})B(x_2)+A(x_{k+1})^2B(x_2)^2+\cdots+(A(x_k)^2B(x_{k+1})^2-2A(x_k)B(x_{k+1})A(x_{k+1})B(x_k)+A(x_{k+1})^2B(x_K)^2$ [using (5)]

$$=\big(A(x_1)B(x_{k+1})-A(x_{k+1})B(x_1)\big)^2+\big(A(x_2)B(x_{k+1})-A(x_{k+1})B(x_2)\big)^2+\cdots+\big(A(x_k)B(x_{k+1})-A(x_{k+1})B(x_k)\big)^2$$

≥ 0, since it is the sum of the square terms.

Hence, by the method of Mathematical Induction (4) is true for all positive integer n and thus (3) also. Therefore, $0 \leq \rho(A, B) \leq 1$. Similarly, it is true for any pair of fuzzy sets A, and B defined over $X \subseteq R$. Thus the formula for finding correlation coefficient is valid for any type of fuzzy sets.

Let us now take numerical examples, to validate our claim. Let X={x1, x2, x3, x4, x5, x6, x7, x8, x9, x10} be a finite set of discrete elements. Let A={(x1, 0.1), (x2, 0.5), (x3, 0.6), (x4, 0.3), (x5, 0), (x6, 0.1), (x7, 0.9), (x8, 1), (x9, 0.8), (x10, 0.3)} and B={(x1, 0.2), (x2, 0.6), (x3, 0.3), (x4, 0.2), (x5, 0.1), (x6, 0.9), (x7, 0.8), (x8, 0.2), (x9, 0.7), (x10, 0.1)} be two fuzzy sets defined over X. Then

$$cov(A,B) = (0.1)(0.2)+(0.5)(0.6)+(0.6)(0.3)+(0.3)(0.2)+(0)(0.1)+(0.1)(0.9)+(0.9)(0.8)+(1)(0.2)+(0.8)(0.7)+(0.3)(0.1)$$

$$= 2.16$$

$$cov(A,A) = (0.1)^2+(0.5)^2+(0.6)^2+(0.3)^2+(0)^2+(0.1)^2+(0.9)^2+(1)^2+(0.8)^2+(0.3)^2$$

=3.6

$$cov(B,B) = (0.2)^2 + (0.6)^2 + (0.3)^2 + (0.2)^2 + (0.1)^2 + (0.9)^2 + (0.8)^2 + (0.2)^2 + (0.7)^2 + (0.1)^2$$

= 2.53

Therefore $\rho(A,B) = \frac{2.16}{\sqrt{(3.6)(2.53)}}$ =0.71572≤1

Obviously $\rho(A,B) \geq 0,\ \rho(A,A) = 1$and $\rho(B,B) = 1$.

Let A={{(x1, 0.2), (x4, 0.4), (x5, 1), (x7, 0.7) , (x10, 0.3)}, and B={(x2, 0.3), (x3, 0.5), (x6, 0.2), (x8, 1), (x9, 0.7)} be two fuzzy sets on X, then

$$cov(A,B) = (0.2)(0) + (0)(0.3) + (0)(0.5) + (0.4)(0) + (1)(0) + (0)(0.2) + (0.7)(0) + (0)(1) + (0)(0.7) + (0.3)(0)$$

=0

$cov(A,A) = (0.2)^2 + (0.4)^2 + (1)^2 + (0.7)^2 + (0.3)^2$=1.78 and $cov(B,B) = (0.3)^2 + (0.5)^2 + (0.2)^2 + (1)^2 + (0.7)^2 = 1.87$

Therefore, $\rho(A,B) = 0$

Again, let X⊆R and A=[2, 4, 6, 9], B=[4, 6, 8, 10] are two trapezoidal number on X, then the membership functions of A and B are given respectively by

$$A(x) = \begin{Bmatrix} 0, & x \leq 2, x \geq 9 \\ \frac{x-2}{2}, & 2 \leq x \leq 4 \\ 1, & 4 \leq x \leq 6 \\ \frac{9-x}{3}, & 6 \leq x \leq 9 \end{Bmatrix} \quad (6)$$

$$B(x) = \begin{Bmatrix} 0, & x \leq 4, x \geq 10 \\ \frac{x-4}{2}, & 4 \leq x \leq 6 \\ 1, & 6 \leq x \leq 8 \\ \frac{10-x}{2}, & 8 \leq x \leq 10 \end{Bmatrix} \quad (7)$$

Now $cov(A,B) = \int_X (A(x))\big(B(x)\big)dx = \int_2^{10} A(x)B(x)dx$

$$= \int_2^4 \frac{(x-2)}{2}(0)dx + \int_4^6 (1)\frac{(x-4)}{2}dx + \int_6^8 \frac{(9-x)}{3}(1)dx + \int_8^9 \frac{(9-x)}{3}\frac{(10-x)}{2}dx + \int_9^{10} (0)\frac{(10-x)}{2}dx$$

= 0.8333333333

$$cov(A,A) = \int_X (A(x)^2\,dx = \int_2^9 A(x)^2 dx$$

$$= \int_2^4 \left(\frac{(x-2)}{2}\right)^2 dx + \int_4^6 (1)^2 dx + \int_6^9 \left(\frac{9-x}{3}\right)^2 dx$$

= 3.6666667

$$cov(B,b) = \int_X (B(x)^2\,dx = \int_4^{10} B(x)^2 dx$$

$$= \int_4^6 \left(\frac{(x-4)}{2}\right)^2 dx + \int_6^8 (1)^2 dx + \int_8^{10} \left(\frac{10-x}{2}\right)^2 dx$$

= 3.5

Therefore, $\rho(A,B) = \frac{0.8333333333}{\sqrt{(3.6666667)(3.5)}} = 0.23262 \leq 1$.

Also $\rho(A,A) = 1, \rho(B,B) = 1$. If A∩B=ϕ (empty set), then $\rho(A,B) = 0$ as $cov(A,B) = 0$.

4. Conclusion

In this article, we have proposed a new definition of correlation coefficient of fuzzy data. We have formulated it in terms of covariance of two fuzzy sets. For this purpose, the covariance of two fuzzy sets is defined using the member functions or values of the fuzzy sets. The formula can be applicable for both discrete and continuous cases. For discrete case, the summation is used and for continuous case integration over the universe of discourse is used. We have also shown that the value of the correlation coefficient lies between 0 and 1. Using numerical examples, we have shown the validity of our method. In future, we will try to use the method in any real system with fuzzy data.

References:

[1] L. A. Zadeh, Fuzzy Sets as Basis of Theory of Possibility, Fuzzy Sets and Systems 1, (1965), pp. 3-28.
[2] M. J. Aczel, and J. Ptanzagl; Remarks on the measurement of subjective probability and information, Metrica, 5, (1966), pp. 91-105.
[3] A. Kandel, and W. J. Byatt, Fuzzy Sets, Fuzzy Algebra and Fuzzy Statistics, Proceedings of the IEEE 66, (1978), pp. 1619-1639.
[4] Teran Pedro, Law of large numbers for possibilistic mean value, Fuzzy Sets and Systems, Vol. 245, (2014), pp. 116-124.
[5] M. J. Son, Correlation of Intuitionistic Fuzzy Sets, Vol 17(4), (2007), pp. 546-549.
[6] T. Gerstenkorn, and J. Manko, Correlation of Intuitionistic fuzzy sets, Fuzzy Sets and Systems, vol. 44, (1991), pp. 29-43.
[7] D. Dumitreascu, A definition of an Informational energy in fuzzy set theory, Sudio Univ, Babes-Bolyat Math 2, (1977), pp. 57-59.
[8] H. Bustince, and P. Burilo, Correlation of Interval-valued intuitionistic fuzzy sets, Fuzzy Sets and Systems, Vol. 74, (1995), pp. 237-244.
[9] R. Saneifard, and R. Saneifard, Correlation Coefficient Between Fuzzy Numbers Based On Central Interval, Journal of Fuzzy Set Valued Analysis, Vol. 2012, (2012), pp. 1-9.
[10] Yu-Ting Cheng, Chih-Ching Yang, The Application of Fuzzy Correlation Coefficient with Fuzzy Interval Data, International Journal of Innovative Management, Information & Production, Volume 5, Number 3, December (2014), pp. 65-71.
[11] G. Hesamiana, and M. Akbarib Ghasem, Testing hypotheses for multivariate normal distribution with fuzzy random variables, International Journal of Systems Science, 2021 Informa UK Limited, trading as Taylor & Francis Group, (2021), pp. 1-11.
[12] J. Klir, and B. Yuan, Fuzzy Sets and Logic Theory and Application, Prentice Hill Pvt. Ltd. (2002).