Fuzzy rank correlation between fuzzy numbers

Pierre-Alexandre Hébert **Marie-Hélène Masson** **Thierry Denœux**

Heudiasyc – UMR CNRS 6599
Université de Technologie de Compiègne
BP 20529 - F-60205 Compiègne Cedex - France

Pierre-Alexandre.Hebert@hds.utc.fr

Abstract

We propose to extend the Kendall rank correlation index τ to intervals and fuzzy numbers. The rank correlation is itself defined as a fuzzy number. An algorithm for computing it is described. A small example illustrates the interest of such a fuzzy index.

Keywords: summary statistics, correlation index, fuzzy ordering relations, Kendall's rank correlation.

1 Introduction

Let $X = \{1, 2, ..., n\}$ be a set of objects. These objects are assumed to be described by two continuous variables U and V. These variables define two strict linear orders L_1 and L_2 on X. A strict linear order on X is a binary relation $L \subseteq X \times X$ that is [8]:

- transitive: $iLj, jLz \Rightarrow iLz \; \forall\, i, j, k \in X$;
- asymmetric[1]: $\forall\, i, j \in X$, not (iLj and jLi);
- complete: $\forall\, i \neq j \in X$, either iLj or jLi.

To evaluate the strength of the relationship between U and V, one can compute Kendall's rank correlation τ which measures the concordance of ranks induced by U and V.

[1]Note that the asymmetry property is equivalent to irreflexivity: $\forall i \in X, not(iLi)$ and antisymmetry: $\forall i, j \in X, iLj$ and $jLi \Rightarrow i = j$.

Let N be the number of pairs of objects: $N = n(n-1)/2$ and D the number of discordant pairs: a pair (i, j) is said *discordant* if it belongs to L_1 and not to L_2, or to L_2 and not to L_1. Then, Kendall's τ is defined as [6]: $\tau = 1 - 2D/N$. Using the set formalism, we get:

$$\tau = \frac{2 \times |L_1 \cap L_2|}{N} - 1$$

In this paper it is proposed to extend this rank correlation coefficient to the case where U and V are not precisely measured, i.e., elements of U and V are intervals or fuzzy numbers. We thus propose a nonparametric counterpart to the fuzzy correlation index introduced in [5]. Sections (2) and (3) present extensions to the case of intervals and fuzzy numbers, respectively. Section (4) describes an example involving fuzzy numbers.

2 Rank correlation between intervals

2.1 Partial ordering relations

Let $[U]$ (respectively $[V]$) be a variable associating an interval $[u_i] = [u_i^-, u_i^+]$ (respectively $[v_i] = [v_i^-, v_i^+]$) to each object $i \in X$.

Let P_1 be the binary relation on X that measures the strict superiority induced by $[U]$ between each pair of objects (i, j): $[u_i] \succ [u_j] \Leftrightarrow iP_1j$. A natural definition of strict superiority of intervals is: $iP_1j \Leftrightarrow u_i^- > u_j^+$. It is trivial to prove that P_1 is transitive and asymmetric: it defines a strict ordering relation. But P_1 is not necessary complete: it is a partial ordering relation.

Let P_2 be the analog partial ordering relation associated to $[V]$. Correlation of $[U]$ and $[V]$ is equivalent to correlation of P_1 and P_2.

2.2 Linear extensions

For a given P, it can be shown that there exists at least one linear order L such that $iPj \Rightarrow iLj$, i.e., $P \subseteq L$. Such a linear order is called a *linear extension* of P. More generally, we associate to each partial order P, a set $E(P)$ containing all *linear extensions* of P.

2.3 Correlation between partial orders

We want to define a concordance measure of two partial orders P_1, P_2. Each partial order P_i can be assimilated to its set of linear extensions $E(P_i)$. Consequently the rank correlation coefficient between P_1 and P_2 may be seen as an interval defined as follows:

$$\tau(P_1, P_2) = [\min_{L_1 \in E(P_1),\, L_2 \in E(P_2)} \tau(L_1, L_2)\ ;\ \max_{L_1 \in E(P_1),\, L_2 \in E(P_2)} \tau(L_1, L_2)]$$

If P_1 and P_2 are linear ordering relations, our definition coincides with the classical Kendall's τ. If not, the size of the interval correlation reflects the lack of information in the partial orders.

2.4 Algorithm

For a given pair of partial orders (P_1, P_2), the problem of finding

$$\tau_{max}(P_1, P_2) = \max_{L_1 \in E(P_1),\, L_2 \in E(P_2)} \tau(L_1, L_2)$$

seems to be difficult. The simplest idea could be to compute $E(P_1)$ and $E(P_2)$, and the corresponding values of τ. However, this approach is computationally prohibitive because of the exponential number of possible linear extensions.

That is the reason why we propose a simple *branch and bound* algorithm that avoids complete generation of linear extensions. Its complexity is still high, and it is not suitable for $n > 10$ objects. In this case we recommend the use of approximations based on some Monte Carlo techniques [1][4].

The algorithm is based on the equivalence of the two following problems:

1. maximizing $\tau(L_1, L_2)$ with L_1, L_2 belonging ,respectively, to $E(P_1)$ and $E(P_2)$;
2. maximizing the cardinal of a binary relation Q (defined on X) such that the transitive closures of $(P_1 \cup Q)$ and $(P_2 \cup Q)$ are asymmetric.

We propose to solve problem (2) by a recursive procedure. We define a function *MaxCard*(P_1, P_2, Q, R) that solves problem (2) by completing a current solution Q, i.e., the optimal relation Q^* must contain all the elements of the third argument Q. We add the constraint that no element of the fourth argument R can participate to Q^*. Indeed R is a binary relation containing all (i, j) that have to be rejected so as to avoid multiple occurrences of a same solution and non-asymmetric solutions. The principle is to test each (i, j) such that $(i, j) \notin Q \cup R$ and such that its insertion does not involve asymmetries by transitive closure. Once a pair (i, j) is accepted, it is combined to the current Q to build a new current solution. The interest of this *branch and bound* approach resides in the asymmetry constraints: the number of admissible pairs (i, j) generally decreases rapidly.

In order to improve the processus, we recommend to add some "*branch and bound* tricks" like minimal and maximal bounds, sorting the admissible pairs of objects (i, j), or better initializations of Q and R.

Here is the description of a basic *MaxCard*. It is supposed that Q is an acceptable solution (i.e., it verifies both constraints of asymmetry).

```
MaxCard(P1, P2, Q, R)
    Score = |Q|
    E = (¬Q) ∩ (¬R)
    For All e ∈ E
        Q' = TransitiveClosure(Q ∪ e)
        T1 = TransitiveClosure(P1 ∪ Q')
        T2 = TransitiveClosure(P2 ∪ Q')
        If T1 and T2 asymmetric Then
            R' = R ∪ T1^-1 ∪ T2^-1
            Score' = MaxCard(P1, P2, Q', R')
            If Score' > Score Then
                Score = Score'
            End If
        End If
        R = R ∪ e
    End For
```

For the first call to *MaxCard*, a simple idea is to take the following values: $Q = P_1 \cup P_2$ and $R = I \cup (P_1)^{-1} \cup (P_2)^{-1}$, with I the identity on X. We are then certain to reach the optimal solution of problem (2) $Score^* = |Q^*|$, and $\tau_{max}(P_1, P_2)$ is then defined by $(2 \times Score^*/N - 1)$.

It is easy to prove that an analog function *MinCard* finding

$$\tau_{min}(P_1, P_2) = \min_{L_1 \in E(P_1),\, L_2 \in E(P_2)} \tau(L_1, L_2)$$

is equivalent to $MaxCard(P_1, P_2^{-1}, Q, R)$.

3 Rank correlation between fuzzy numbers

3.1 Fuzzy partial ordering relations

A fuzzy relation $\widetilde{P} \in [0,1]^{X \times X}$ is a fuzzy strict partial order iff it verifies the following properties (we note $\widetilde{P}(i,j)$ instead of $\mu_{\widetilde{P}}(i,j)$) [8]:

- max-min transitivity:
 $\forall\, i,j,k \in X,\ \widetilde{P}(i,k) \geq \bigvee_j \widetilde{P}(i,j) \wedge \widetilde{P}(j,k)$;
- asymmetry: $\forall\, i,j \in X,\ \widetilde{P}(i,j) \wedge \widetilde{P}(j,i) = 0$.

Any α-cut $\widetilde{P}_\alpha$ of a fuzzy partial order is a crisp partial order. For all $0 < \alpha_1 < \alpha_2 \leq 1$, $\widetilde{P}_{\alpha_1} \supseteq \widetilde{P}_{\alpha_2}$: the α-cuts $\widetilde{P}_\alpha$ form a nested sequence of distinct partial orders.

Let $\widetilde{U}$ (respectively $\widetilde{V}$) be a variable associating a fuzzy number $\widetilde{u}_i$ (respectively $\widetilde{v}_i$) to each object $i \in X$.

In Section 2.1, we used a generalized measure of superiority of crisp intervals; we now define such a measure for fuzzy numbers. To compare fuzzy numbers, the following measure is generally used [2]:

$$\widetilde{R}(i,j) = \sup_{a \geq b} \min(\mu_{\widetilde{u}_i}(a), \mu_{\widetilde{u}_j}(b))$$

Such a relation is not transitive, but its dual $\widetilde{P}_1(i,j) = 1 - \widetilde{R}(j,i)$ is a measure of the necessity of $\widetilde{u}_i \geq \widetilde{u}_j\, \forall (i,j)$, and it can be shown that it defines a fuzzy strict partial ordering relation [7]: $\widetilde{P}_1(i,j)$ indicates in which degree $\widetilde{u}_i$ is strictly superior to $\widetilde{u}_j$.

Let $\widetilde{P}_2$ be the analog fuzzy partial relation associated to $\widetilde{V}$. Correlation of $\widetilde{U}$ and $\widetilde{V}$ is equivalent to correlation of $\widetilde{P}_1$ and $\widetilde{P}_2$.

3.2 Correlation between fuzzy orders

We can associate to a partial order P the set $E(P)$ of its linear extensions. We want to extend this notion to a fuzzy order $\widetilde{P}$. Let $\mathcal{L}(X) \subseteq 2^{X \times X}$ denote the set of linear orders on X. The fuzzy set $E(\widetilde{P}) \in [0,1]^{\mathcal{L}(X)}$ of linear orderings of $\widetilde{P}$ may be defined by its membership function:

$$\mu_{E(\widetilde{P})}(L) = I(\widetilde{P}, L) \quad \forall L \in \mathcal{L}(X)$$

where I is an inclusion index. Using a common choice [3] we arrive at:

$$\begin{aligned} \mu_{E(\widetilde{P})}(L) &= 1 - ht(\widetilde{P} \cap \overline{L}) \\ &= 1 - \sup_{i,j} \min(\widetilde{P}(i,j), 1 - L(i,j)) \end{aligned}$$

Let $\alpha \in (0,1]$. The α-cut of $E(\widetilde{P})$ is the crisp set $E(\widetilde{P})^\alpha$ defined by all $L \in \mathcal{L}(X)$ that verify:

$$1 - \sup_{i,j} \min(\widetilde{P}(i,j), 1 - L(i,j))) \geq \alpha$$

We thus have:

$$\begin{aligned} & L \in E(\widetilde{P})^\alpha \\ \Leftrightarrow\ & \sup_{i,j} \min(\widetilde{P}(i,j), 1 - L(i,j))) \leq 1 - \alpha \\ \Leftrightarrow\ & \forall (i,j), \min(\widetilde{P}(i,j), 1 - L(i,j))) \leq 1 - \alpha \\ \Leftrightarrow\ & \forall (i,j), \widetilde{P}(i,j) > 1 - \alpha \Rightarrow 1 - L(i,j) = 0 \\ \Leftrightarrow\ & \forall (i,j) \in \widetilde{P}^{(1-\alpha)+},\ (i,j) \in L \\ \Leftrightarrow\ & L \in E(\widetilde{P}^{(1-\alpha)+}) \end{aligned}$$

We can no extend the notion of partial ordering correlation interval $\tau(P_1, P_2)$ to fuzzy partial ordering relations: $\tau(\widetilde{P}_1, \widetilde{P}_2)$ becomes a fuzzy interval. It is defined by its α-cut in the following way:

$$\tau(\widetilde{P}_1, \widetilde{P}_2)^\alpha = \tau(P_1^{(1-\alpha)+}, P_2^{(1-\alpha)+})$$

4 Example

Let X be a set of $n = 9$ students, that are scored on $c = 5$ distinct courses: respectively *MAThematics, PHYsics, LITerature, LATin, DRAwing.*

Scores are either precise or imprecise, and are represented by fuzzy trapezoidal numbers as shown in Fig. 1.

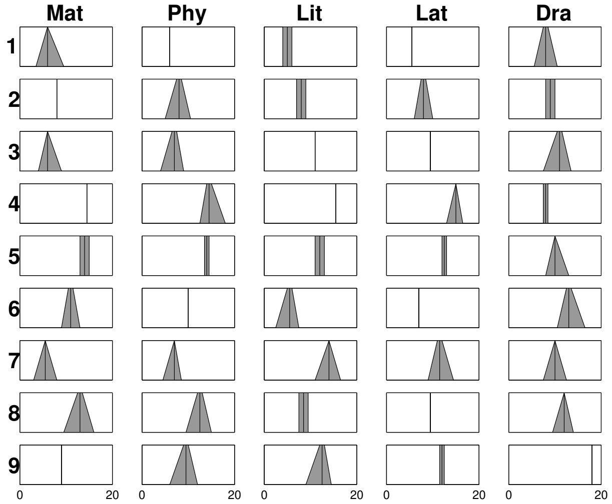

Figure 1: Fuzzy scores (students × courses)

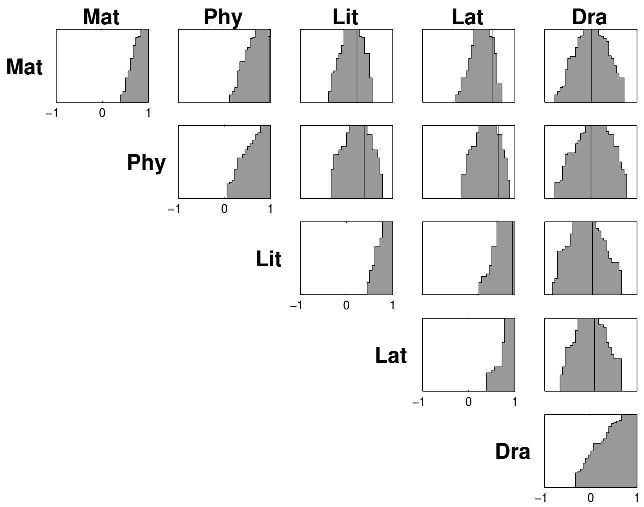

Figure 2: Fuzzy correlations (courses × courses)

Results given in Fig. 2 confirm the non-surprising correlation between MATHematics and PHYsics: "masses" are clearly on the right hand side. Between these two courses and LATin, one can detect a weak correlation. Conversely, LATin and LITerature have a very high correlation. Other correlations do not seem relevant. On the diagonal, the fuzzy auto-correlations indicate how far fuzzy partial orders are from linear orders.

Vertical lines indicate the values of Kendall's τ obtained by replacing fuzzy intervals by the middles of their cores. Nonfuzzy correlations between Latin and both PHYsics and MATHematics clearly look optimistic: little changes result in drastic transformations in ordering relations.

5 Conclusion

An extension of rank correlations to fuzzy orders and fuzzy numbers has been proposed. Potential applications are multicriteria decision making (to compare fuzzy orderings of alternatives according to different criteria or different decision makers), and nonparametric statistical analysis of imprecise data.

References

[1] R. Bubley and D.Dyer. Faster random generation of linear extensions. In *Proceedings of the Ninth Annual ACM-SIAM Symposyum on Discrete Algorithms*, pages 175-186, San Francisco, CA, 1998.

[2] D. Dubois and H. Prade. Possibility Theory. An Approach to Computerized Processing of Uncertainty, Plenum Press, New York, 1988.

[3] D. Dubois and H. Prade. Fuzzy sets: history and basic notions. In D. Dubois and H. Prade (Eds), *Fundamentals of fuzzy sets*, chap. 4, 21–124. Kluwer, Boston, 2000.

[4] V. Ha and P. Haddawy. Similarity of personal preferences: Theoretical foundations and empirical analysis. *Artificial Intelligence*, 146 (2):149-173, 2003.

[5] S.-T. Liu and C. Kao. Fuzzy measures for correlation coefficient of fuzzy numbers. *Fuzzy Sets and Systems*, 128 (2):267-275, 2002.

[6] Kendall M.G. Rank Correlation Methods. Hafner, New York, 1970.

[7] S. Ovchinnikov. An introduction to fuzzy relations. In D. Dubois and H. Prade (Eds), *Fundamentals of fuzzy sets*, chap. 4, 233–259. Kluwer, Boston, 2000.

[8] L.A. Zadeh. Similarity Relations and Fuzzy Orderings. *Information Sciences* (3): 177-200, 1971.