Kendall's Rank Correlation on Quantized Data: An Interval-Valued Approach

Inés Couso[a], Olivier Strauss[b], Hugo Saulnier

[a]*Dep. of Statistics and O.R., University of Oviedo, Spain*
[b]*LIRMM, Montpellier University, France*

Abstract

Kendall's rank correlation coefficient, also called Kendall's τ, is an efficient and robust way for identifying monotone relationships between two data sequences. However, when applied to digital data, the high number of ties yields inconsistent results due to quantization. Here, we propose an extension of Kendall's τ that considers an epistemic view of a sequence of quantized data − each sample is supposed to be the quantized version of an original value that is a real number. We come up with an imprecise τ, defined as the interval containing all τ values that could have been computed on sequences of original values before quantization. We propose a very simple and straightforward algorithm to compute this interval-valued τ. We prove the exactness of the bounds and propose an experiment that illustrates the need for such an extension.

Keywords: Correlation, ties, interval-valued statistics

1. Introduction

1.1. Correlation in Image Processing

Correlation coefficients are very well-known statistical tools that are extensively used in statistical analysis, pattern recognition, and image processing. Applications for the latter include comparing two images for the purpose of image registration, object recognition, disparity measurement, motion estimation, etc., with applications ranging from pattern recognition, surveillance or authentication to video analysis.

The simplest and perhaps oldest correlation coefficient is the Pearson correlation coefficient [12]. It is simply the scalar product between the mean-adjusted and normalized versions of the two concerned sequences. However, the Pearson correlation coefficient only measures linear dependence relations. In image processing, this linearity hypothesis is usually non-valid because of a variety of factors, such as intensity distortion due to camera gain, offset or saturation [4]. Moreover, this type of correlation is known to be very sensitive to non-stationary noise and outlier values [11].

On the other hand, Kendall's correlation coefficient, or simply called Kendall's τ

(or just τ), is more suited for use in image processing where stationarity cannot usually be advocated [9]. Computation of τ involves comparing the ranking of pairs of samples in the two sequences. It allows us to identify not only linear dependencies between the sequences, but also any kind of monotonic relationship. Moreover, being a rank statistic, it is more robust to outliers and non-parametric hypothesis tests can be derived from it [10].

Like other correlation coefficients, Kendall's τ was designed to deal with real-valued datasets, where ties are uncommon. Yet, pixel values in image processing are digital, i.e. quantized, values. Thus a set of pixel values is likely to contain many ties. Very few studies have focused on the effect of data quantization on Kendall's correlation coefficient, while using a statistical tool designed for real data for analyzing quantized data can lead to wrong conclusions [13].

1.2. Definition of Kendall's τ and variants

Kendall's τ is defined as the difference between the number of pairs of *concordant* values and of pairs of *discordant* values, normalized by the total number of pairs. Let $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ be two sequences to compare. Let A be the set of all pairs of indices, i.e.,

$$A = \{(i, j) \in \{1, \dots, n\} \times \{1, \dots, n\} \mid i < j\},$$

and let $n_A = \frac{1}{2}n(n-1)$ denote its cardinal. Let $C \subseteq A$, with cardinal n_C, be the subset of concordant pairs and $D \subseteq A$, with cardinal n_D, be the subset of discordant pairs, with both defined as:

$$C = \{(i, j) \in A \mid (x_i < x_j \text{ and } \ y_i < y_j) \text{ or } (x_i > x_j \text{ and } y_i > y_j)\},$$

$$D = \{(i, j) \in A \mid (x_i < x_j \text{ and } \ y_i > y_j) \text{ or } (x_i > x_j \text{ and } y_i < y_j)\},$$

Kendall's rank correlation is then defined as:

$$\tau(\mathbf{x}, \mathbf{y}) = \frac{n_C - n_D}{n_A}.$$

This coefficient ranges from 1 when $n_C = n_A$ (only concordant pairs) to -1 when $n_D = n_A$ (only discordant pairs).

One major drawback of this formulation is that pairs which contain an equality (either $x_i = x_j$ or $y_i = y_j$) are included in neither C nor D. Hence, $n_A > n_D + n_C$ and the bounds of τ can never be reached, thus flawing the interpretation of the coefficient. Two variants of τ can be considered to deal with this problem: γ and τ_b [15]. They both consist of normalizing the coefficient so that proper bounds can be reached.

The variant γ consists of normalizing the coefficient by $n_C + n_D$ rather than by the total number of pairs.

$$\gamma(\mathbf{x}, \mathbf{y}) = \frac{n_C - n_D}{n_C + n_D}.$$

Within this variant, pairs which contain an equality are considered as outliers.

The variant τ_b needs to consider two new subsets of A, U and V, which respectively contain pairs with an equality in the first sequence and those in the second:

$$U = \{(i,j) \in A \mid x_i = x_j\},$$
$$V = \{(i,j) \in A \mid y_i = y_j\},$$

The variant τ_b is then defined as:

$$\tau_b(\mathbf{x}, \mathbf{y}) = \frac{n_C - n_D}{\sqrt{n_A - n_U}\sqrt{n_A - n_V}},$$

where n_U (rsp. n_V) is the cardinal of U (rsp. V).

Remark 1. *τ and its variants τ_b and γ are always ordered in the following way: $\tau \leq \tau_b \leq \gamma$.*

Proof: Let $T \subseteq A$, with cardinal n_T, be the subset of ties:

$$T = \{(i,j) \in A \mid x_i = x_j \text{ or } y_i = y_j\}.$$

$n_C + n_D = n_A - n_T$, therefore $n_C + n_D < n_A$ and thus $\tau \leq \gamma$. By definition $U \subseteq T$, thus $n_A - n_U \geq n_A - n_T$, and $V \subseteq T$, thus $n_A - n_V \geq n_A - n_T$. Therefore $\tau_b \leq \gamma$. Naturally, since $n_A \geq n_A - n_U$ and $n_A \geq n_A - n_V$, $\tau \leq \tau_b$. $\square$

1.3. Adding imprecision in the Kendall's correlation coefficient: previous work

In the relevant literature, few authors have considered including imprecision when computing the Kendall's correlation coefficient. Two main pathways have been explored. On one hand, in [1], and then in [14], the problem of a lack of robustness of γ with respect to observation noise when the two observables have ties is addressed. The new Kendall's coefficient they propose is based on a (fuzzy) set-valued modification of the ranking. On the other hand, in [3], the problem of computing the Kendall's correlation coefficient when the observations are imprecise is addressed. This problem is closer to that addressed in this paper.

1.3.1. Fuzzy gamma rank correlation coefficient

Let us consider that the two sequences $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ have many ties, i.e. n_T cannot be neglected. Let us suppose that those sequences have been observed by a noisy sensor providing the observed sequences $\mathbf{x^o} = (x_1^o, \ldots, x_n^o)$ and $\mathbf{y^o} = (y_1^o, \ldots, y_n^o)$. Because of the noise, it is likely that many tied pairs (i,j) that belong to T, the set of ties in the original data, does not belong to T^o, the set of ties in the observed data. Hence, γ^o, the γ-variant of τ based on the observed sequences, is likely to be a biased estimation of the true γ, based on the original sequence. This phenomenon has been thoroughly analyzed in [14]. These authors show that the observed concordant (rsp. discordant) pairs C^o (rsp. D^o) is likely included in the original concordant (rsp. discordant) pairs C (rsp. D). The robustness of the Kendall's correlation coefficient towards outliers and the non-linearity of the relation between the two sequences has the drawback of a certain degree of sensitivity towards small changes. Replacing the strict ordering, used in the original Kendall's correlation coefficient, by a fuzzy order relation leads to less sensitivity of the obtained rank correlation. This reduced sensitivity is due to the lower influence of negligible changes in the observed data, that could likely be explained by the effect of noise in the ties.

The problem addressed in the above cited papers has little to do with the problem we address here, since quantized data cannot be considered as noisy measurements of the original data. Moreover, quantized data are more likely to have ties than the original sequence.

1.3.2. Kendall's correlation coefficient for imprecise data

The problem considered here is closer to the one we address in this paper. Let us suppose that the observation process leads to interval-valued observed data $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$, where X_i (rsp. Y_i) is an interval-valued observation of x_i (rsp. y_i). The most general formalization, proposed by Denœux et al., consists of finding the lower bound (τ^-) and the upper bound (τ^+) of the set of all Kendall's τ that could have been obtained by considering any pair of sequences $\mathbf{x} \in \mathbf{X}$ and $\mathbf{y} \in \mathbf{Y}$. Their approach is to consider a partial order over a set of intervals and minimize (resp. maximize) τ. Then there is simply a problem of finding which linear extension (i.e. complete ordering compatible with the partial ordering) yields the minimal (resp. maximal) value. But since the number of linear extensions can possibly grow exponentially with the size of the dataset, an exhaustive search on the space of linear extensions is not very feasible because of the high computational expense. They propose a first alternative in [3] with a *branch and bound* algorithm which proceeds recursively. This algorithm is certain to reach the optimal bounds, but its complexity makes it impractical to use when the data exceeds 10 objects. They later [6] refined their approach by applying Monte-Carlo sampling directly on the space of all linear extensions. The choice of the Bubbley and Dyer algorithm ensures uniform sampling on the space and sufficient computational efficiency for 30 objects to be processed at most. As a main drawback, this approach leads to an overestimation τ_* (rsp. underestimation τ^*) of the lower (rsp. upper) bound of the sought after interval $[\tau^-, \tau^+]$.

Hryniewicz and Szediw investigated Kendall's correlation for interval data in the case of autocorrelated data [8]. They proposed a heuristic algorithm to find an approximation of bounds which has proven efficient on larger datasets (≥ 50). However, it offers limited generalization, since the optimization method relies on the fact that the compared sequences of interval data are actually a single sequence and its time-shifted version. An algorithm for the general case was then proposed in [7]. Here the idea is to exploit *patterns* which are known to yield maximal or minimal τ results. After sorting the data to be as close as possible to one of these *patterns*, an iteration procedure brings the algorithm closer to an optimal bound. Although computationally efficient, this algorithm gives varied results depending on the choice of heuristic.

The problem we address in this paper is more specific. As we will show below, quantized values can naturally be considered as imprecise observations of real values but with two main restrictions that make the computational problem more tractable: 1- imprecise observations can only take a limited number of values (quantized values), 2- those values form an ordered partition of the whole range. The solution we propose also leads to an interval-valued estimation of the correlation coefficient but the bounds are exact and easy to compute.

1.4. The quantization problem

Since the rise of digital technologies, quantized data has become ubiquitous. Inherently, any sequence of quantized data contains ties and the lower the number of quantization levels, the higher the number of ties. In image processing, for example, usual gray-scale images are quantized on 8 bits, hence on 256 different values. Even for small-scale 256 × 256 pixel images, an average of 256 pixels will share the same value. Saturations are also likely to produce many ties.

As noted previously, variants (τ_b and γ) have been introduced in order to make Kendall's correlation useable on data containing ties. However, we might object to the fact that those two variants merely consist of removing the ties of the analysis. This means that these samples are considered as *outliers* whereas they are actually just data that do not fit in a concordant versus discordant predicate. More exactly, their classification is unknown or imprecise. In image processing, as well as in many other fields, it is more than likely that the original sequences (before quantization) contain little or even no ties. Simply removing samples that cannot be classified means that part of the information is not considered in computing the correlation. Thus all further reasoning based on this correlation coefficient would be unsound. The question then becomes, "How can all information contained in the quantized data be meaningfully utilized?"

2. Imprecisely valued τ

2.1. Quantization modelled as intervals

The imprecision introduced by quantization can easily be represented by an interval. We here adopt an epistemic [2] point of view: a quantization interval represents the sequence of values that a piece of information could have taken before quantization. Let p and q be the quantization levels for X and Y, i.e. quantized values belonging to $\{1, \ldots p\}$ and $\{1, \ldots, q\}$, respectively. Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ be the original real-valued sequences (before quantization). Let $\Phi_k^1 = [\underline{\phi_k^1}, \overline{\phi_k^1}[$ be the real-valued interval associated with the k^{th} quantization level, $k \in \{1, \ldots, p\}$ of the first variable. Analogously, let $\Phi_l^2 = [\underline{\phi_l^2}, \overline{\phi_l^2}[$ be the interval associated with the l-th quantization level of the second variable, for $l \in \{1, \ldots, q\}$. Inherently $\Phi_k^1 \cap \Phi_{k'}^1 = \emptyset$ and $\Phi_l^2 \cap \Phi_{l'}^2 = \emptyset$ as well, if $k \neq k'$ and $l \neq l'$. Let $\mathbf{x}^Q = (x_1^Q, \ldots, x_n^Q)$ and $\mathbf{y}^Q = (y_1^Q, \ldots, y_n^Q)$ denote the quantized sequences. Indeed, those x_i^Q and y_i^Q are the only samples we have access to. By construction, $\forall i \in \{1, \dots n\}$, $x_i^Q = k$ (resp. $y_i^Q = l$) is equivalent to $x_i \in \Phi_k^1$ (resp. to $y_i \in \Phi_l^2$). Let $\boldsymbol{\Phi}^1$ and $\boldsymbol{\Phi}^2$ denote the collections of intervals associated with the p quantized values of the first variable and the q quantized values of the second one, i.e. $\boldsymbol{\Phi}^1 = \{\Phi_1^1, \dots, \Phi_p^1\}$ and $\boldsymbol{\Phi}^2 = \{\Phi_1^2, \dots, \Phi_q^2\}$. Thus to each pair of quantized values (x_i^Q, y_i^Q) we can associate the pair of intervals $X_i = [\underline{x}_i, \overline{x}_i[\; \in \boldsymbol{\Phi}^1$ and $Y_i = [\underline{y}_i, \overline{y}_i[\; \in \boldsymbol{\Phi}^2$. Those intervals represent our incomplete knowledge about the hypothetic original *true* values x_i and y_i. Therefore, within this epistemic view, dealing with a quantized valued dataset is equivalent to dealing with a set of disjoint intervals.

Now consider a quantized vector $(x_1^Q, \dots, x_n^Q) \in \{1, \dots, l\}^n$, and let the two sequences of intervals $\mathbf{X} = (X_1, \dots, X_n)$ and $\mathbf{Y} = (Y_1, \dots, Y_n)$ represent the collection of all possible sequences $\mathbf{x}$ and $\mathbf{y}$ before quantization. We can identify three mutually exclusive cases for any pair $(i, j) \in A$ with respect to $\mathbf{X}$:

(a) $x_i^Q < x_j^Q \Leftrightarrow \overline{x_i} \leq \underline{x_j}$

(b) $x_i^Q > x_j^Q \Leftrightarrow \underline{x_i} \geq \overline{x_j}$

(c) $x_i^Q = x_j^Q \Leftrightarrow (\overline{x_i} \geq \underline{x_j}$ and $\underline{x_i} \leq \overline{x_j})$

The above equations are closely related to the "interval dominance" concept [5]:

$$[a, b[\; \leq_{ID} [c, d[\text{ iff } b \leq c.$$

The relation $\leq_{ID}$ is a partial ordering over the set of bounded intervals of the real line. The above equations can be equivalently written as follows:

$$\begin{aligned} x_i^Q < x_j^Q &\Leftrightarrow X_i \leq_{ID} X_j \\ x_i^Q > x_j^Q &\Leftrightarrow X_i \geq_{ID} X_j \\ x_i^Q = x_j^Q &\Leftrightarrow (X_i \not\leq_{ID} X_j \text{ and } X_i \not\geq_{ID} X_j) \end{aligned} \tag{1}$$

Interval dominance indicates that two different intervals are incomparable as soon as they overlap. But let us note that, given the nature of quantization, the compared intervals are either coincident ($X_i = X_j$) or disjoint. Furthermore, the collection $\boldsymbol{\Phi}^i$ $(i = 1, 2)$ does not contain any singleton and therefore the restriction of $\leq_{ID}$ to $\boldsymbol{\Phi}^i$ is a strict total ordering that we can denote $<_{ID}$. Therefore, we can write:

- $x_i^Q < x_j^Q \Leftrightarrow X_i <_{ID} X_j$
- $x_i^Q > x_j^Q \Leftrightarrow X_i <_{ID} X_j$
- $x_i^Q = x_j^Q \Leftrightarrow X_i = X_j$, and similarly
- $y_i^Q < y_j^Q \Leftrightarrow Y_i <_{ID} Y_j$
- $y_i^Q > y_j^Q \Leftrightarrow Y_i <_{ID} Y_j$

- $y_i^Q = y_j^Q \Leftrightarrow Y_i = Y_j$.

We can easily derive the following result:

Lemma 1. *The collection of subsets of* A, $\{S_1^Q, L_1^Q, U^Q\}$ *defined as follows:*

- $S_1^Q = \{(i,j) \in A \,|\, X_i <_{ID} X_j\} = \{(i,j) \in A : x_i^Q < x_j^Q\}$,
- $L_1^Q = \{(i,j) \in A \,|\, X_i >_{ID} X_j\} = \{(i,j) \in A : x_i^Q > x_j^Q\}$,
- $U^Q = \{(i,j) \in A \,|\, X_i = X_j\} = \{(i,j) \in A : x_i^Q = x_j^Q\}$,

forms a partition of A. *Analogously, the collection* $\{S_2^Q, L_2^Q, V^Q\}$ *defined as:*

- $S_2^Q = \{(i,j) \in A \,|\, Y_i <_{ID} Y_j\} = \{(i,j) \in A : y_i^Q < y_j^Q\}$,
- $L_2^Q = \{(i,j) \in A \,|\, Y_i >_{ID} Y_j\} = \{(i,j) \in A : y_i^Q > y_j^Q\}$,
- $V^Q = \{(i,j) \in A \,|\, Y_i = Y_j\} = \{(i,j) \in A : y_i^Q = y_j^Q\}$.

also forms a partition of A.

2.2. Optimal bounds of imprecise τ

Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ be two sequences and suppose that we only have access to a pair of sequences of intervals $(\mathbf{X}, \mathbf{Y})$ representing our incomplete information about them. Let $\mathbb{E}(\mathbf{X}, \mathbf{Y})$ denote the set of all of those feasible pairs of samples, i.e.:

$$\mathbb{E}(X, Y) = \{(\mathbf{x}', \mathbf{y}') \,|\, (x_i', y_i') \in X_i \times Y_i,\ i = 1, \ldots, n\}.$$

According to this partial information about the initial pair of sequences, their correlation coefficients τ, γ and τ_b can be any value inside the respective sets:

$$\tau(\mathbb{E}(\mathbf{X}, \mathbf{Y})) = \{\tau(\mathbf{x}', \mathbf{y}') \,|\, (\mathbf{x}', \mathbf{y}') \in \mathbb{E}(\mathbf{X}, \mathbf{Y})\}.$$

$$\gamma(\mathbb{E}(\mathbf{X}, \mathbf{Y})) = \{\gamma(\mathbf{x}', \mathbf{y}') \,|\, (\mathbf{x}', \mathbf{y}') \in \mathbb{E}(\mathbf{X}, \mathbf{Y})\}.$$

$$\tau_b(\mathbb{E}(\mathbf{X}, \mathbf{Y})) = \{\tau_b(\mathbf{x}', \mathbf{y}') \,|\, (\mathbf{x}', \mathbf{y}') \in \mathbb{E}(\mathbf{X}, \mathbf{Y})\}.$$

We will call the above set the *set-valued extension* of τ, γ and τ_b respectively. In this setting, $\tau^+(\mathbf{X}, \mathbf{Y})$ and $\tau^-(\mathbf{X}, \mathbf{Y})$ respectively denote the maximum and minimum of the set $\tau(\mathbb{E}(\mathbf{X}, \mathbf{Y}))$.

Let us now consider the following subsets of $A = \{(i,j) \in \{1, \ldots, n\} \times \{1, \ldots, n\} \,|\, i < j\}$:

- The set of pairs that are *necessarily concordant*,

$$C^Q = \{(i,j) \in A \,|\, [X_i <_{ID} X_j \text{ and } Y_i <_{ID} Y_j], \text{ or } [X_i >_{ID} X_j \text{ and } Y_i >_{ID} Y_j]\}. \quad (2)$$

- The set of pairs that are *necessarily discordant*,

$$D^Q = \{(i,j) \in A \,|\, [X_i >_{ID} X_j \text{ and } Y_i <_{ID} Y_j], \text{ or } [X_i <_{ID} X_j \text{ and } Y_i >_{ID} Y_j]\}. \tag{3}$$

- The set of pairs that are *unordered*, or neither *necessarily concordant* nor *necessarily discordant.*

$$T^Q = \{(i,j) \in A \mid X_i = X_j \text{ or } Y_i = Y_j\} = U^Q \cup V^Q. \tag{4}$$

Definition 1. *We define the* interval-valued Kendall's τ*,* $\underline{\overline{\tau}}(\mathbf{x}^Q, \mathbf{y}^Q) = [\underline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q), \overline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q)]$*, as:*

$$\underline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q) = \frac{n_{C^Q} - n_{D^Q} - n_{T^Q}}{n_A},$$

$$\overline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q) = \frac{n_{C^Q} - n_{D^Q} + n_{T^Q}}{n_A}.$$

The main goal in this section is to prove that the extremes of the interval-valued τ do respectively coincide with the minimum and maximum values of the sets $\tau(\mathbb{E}(\mathbf{X}, \mathbf{Y}))$, $\gamma(\mathbb{E}(\mathbf{X}, \mathbf{Y}))$ and $\tau_b(\mathbb{E}(\mathbf{X}, \mathbf{Y}))$, i.e. $\overline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q) = \tau^+(\mathbf{X}, \mathbf{Y})$ and $\underline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q) = \tau^-(\mathbf{X}, \mathbf{Y})$.

We will solve our problem in two phases. We will first prove that those extreme points bound the above sets. In a second step, we will prove that those extreme points are indeed included in the three of them.

2.2.1. Bounds for set-valued extension of Kendall's τ and its variants

In this subsection, we will prove that the three sets $\tau(\mathbb{E}(\mathbf{X}, \mathbf{Y}))$, $\tau_b(\mathbb{E}(\mathbf{X}, \mathbf{Y}))$ and $\gamma(\mathbb{E}(\mathbf{X}, \mathbf{Y}))$ are lower (resp. upper) bounded by $\underline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q)$ (resp. by $\overline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q)$).

Let us consider an arbitrary paired sample $(\mathbf{x}, \mathbf{y}) \in \mathbb{E}(\mathbf{X}, \mathbf{Y})$ and let C and D respectively denote the collection of concordant and discordant pairs in it. Let $T = A \setminus (C \cup D) = U \cup V$ denote the rest of the pairs (pairs with a tie in at least one of the components). Then we can easily check the following result:

Lemma 2. *The following relations hold:*

- $C^Q \subseteq C$.
- $D^Q \subseteq D$.

Proof: Let us first remind the reader that, given an arbitrary pair $(i,j) \in A$, $X_i = [\underline{x_i}, \overline{x_i}[$ is interval dominated by $X_j = [\underline{x_j}, \overline{x_j}[$ (i.e. $X_i <_{ID} X_j$) if and only if $\overline{x_i} \leq \underline{x_j}$. Therefore, regardless of the specific values of $x_i \in X_i = [\underline{x_i}, \overline{x_i}[$ and $x_j \in X_j = [\underline{x_j}, \overline{x_j}[$, the relation $X_i <_{ID} X_j$ implies the strict inequality $x_i < x_j$. Analogously, $Y_i <_{ID} Y_j$ implies that $y_i < y_j$, for every pair $(y_i, y_j) \in Y_i \times Y_j$. From the above implications we deduce that the set of necessarily concordant pairs:

$$C^Q = \{(i,j) \in A \mid [X_i <_{ID} X_j \text{ and } Y_i <_{ID} Y_j] \text{ or } [X_i >_{ID} X_j \text{ and } Y_i >_{ID} Y_j]\}$$

is included in the set of concordant pairs

$$C = \{(i,j) \in A \,|[x_i < x_j \text{ and } y_i < y_j] \text{ or } [x_i > x_j \text{ and } y_i > y_j]\}.$$

Lemma 3. *$\{C^Q, D^Q, T^Q\}$ forms a partition of A.*

Proof: The sets C^Q, D^Q and T^Q can be respectively expressed as follows in terms of the notation introduced in Lemma 1:

$$\begin{aligned} C^Q &= (S_1^Q \cap S_2^Q) \cup (L_1^Q \cap L_2^Q), \\ D^Q &= (L_1^Q \cap S_2^Q) \cup (S_1^Q \cap L_2^Q), \\ T^Q &= U^Q \cup V^Q. \end{aligned} \tag{5}$$

According to the distributive property, we observe that:

- $C^Q \cap D^Q \subseteq (S_1^Q \cap L_1^Q) \cup (S_2^Q \cap L_2^Q)$.
- $C^Q \cup D^Q = (S_1^Q \cup L_1^Q) \cap (S_2^Q \cup L_2^Q)$.

Furthermore, according to Lemma 1, the collections $\{L_1^Q, S_1^Q, U^Q\}$ and $\{L_2^Q, S_2^Q, V^Q\}$ respectively partition the set A. Therefore, we have: $S_1^Q \cap L_1^Q = S_2^Q \cap L_2^Q = \emptyset$, $S_1^Q \cup L_1^Q = (U^Q)^c$ and $S_2^Q \cup L_2^Q = (V^Q)^c$. Thus, we deduce that

- $C^Q \cap D^Q = \emptyset$
- $C^Q \cup D^Q = (U^Q)^c \cap (V^Q)^c = (U^Q \cup V^Q)^c = (T^Q)^c$,

and we easily derive that $\{C^Q, D^Q, T^Q\}$ forms a partition of A.

Lemma 4. *The following relations hold:*

(a) $C \cup U \cup V \subseteq C^Q \cup T^Q$,

(b) $D \cup U \cup V \subseteq D^Q \cup T^Q$,

(c) $U \cup V \subseteq T^Q$.

Proof: The above three statements can be derived from the following facts:

- By virtue of Lemma 3, $\{C^Q, D^Q, T^Q\}$ forms a partition of A, and therefore $C^Q \cup T^Q = (D^Q)^c$, $D^Q \cup T^Q = (C^Q)^c$ and $T^Q = (C^Q \cup D^Q)^c$.
- Analogously, $\{C, D, U, V\}$ forms a partition of A and therefore $D \cup U \cup V = C^c$, $C \cup U \cup V = D^c$ and $U \cup V = (C \cup D)^c$.
- According to Lemma 2, $C^Q \subseteq C$ and $D^Q \subseteq D$, and therefore, $C^c \subseteq (C^Q)^c$, $D^c \subseteq (D^Q)^c$ and $(C \cup D)^c \subseteq (C^Q \cup D^Q)^c$.

We need an additional auxiliary lemma in order to prove the main result in this subsection.

Lemma 5. *Consider three real numbers a, b, x satisfying the following restrictions:*

- $a \leq b$*,* $b > 0$*.*
- $x \geq 0$*.*

Then the following inequality holds

$$\frac{a+x}{b+x} \geq \frac{a}{b}.$$

Proof: Taking into account the above hypotheses, we deduce that $0 < b+x$ and therefore:

$$\frac{a+x}{b+x} \geq \frac{a}{b} \Leftrightarrow (a+x)b \geq (b+x)a \Leftrightarrow x \cdot b \geq x \cdot a \Leftrightarrow x\,(b-a) \geq 0,$$

which is trivially satisfied, according to the hypotheses.

Let us now prove the main result in this subsection.

Theorem 1. $\underline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q)$ *and* $\overline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q)$ *bound the three sets* $\tau(\mathbb{E}(\mathbf{X}, \mathbf{Y}))$*,* $\gamma(\mathbb{E}(\mathbf{X}, \mathbf{Y}))$ *and* $\gamma(\mathbb{E}(\mathbf{X}, \mathbf{Y}))$ *respectively from below and above.*

Proof: According to Remark 1, given an arbitrary sample $(\mathbf{x}, \mathbf{y}) \in \mathbb{E}(\mathbf{X}, \mathbf{Y})$, the following inequalities hold: $\tau \leq \tau_b \leq \gamma$. Therefore, we simply have to prove that $\underline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q) \leq \tau(\mathbf{x}, \mathbf{y})$ and that $\gamma(\mathbf{x}, \mathbf{y}) \leq \overline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q)$, for every $(\mathbf{x}, \mathbf{y}) \in \mathbb{E}(\mathbf{X}, \mathbf{Y})$.

- Let us first prove that $\underline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q) \leq \tau(\mathbf{x}, \mathbf{y})$, for every $(\mathbf{x}, \mathbf{y}) \in \mathbb{E}(\mathbf{X}, \mathbf{Y})$. We just have to show that $n_{C^Q} - (n_{T^Q} + n_{D^Q}) \leq n_C - n_D$, where n_C and n_D respectively denote the number of concordant and discordant pairs in an arbitrary sample $(\mathbf{x}, \mathbf{y}) \in \mathbb{E}(\mathbf{X}, \mathbf{Y})$. The above inequality is implied by the pair of inequalities $n_{C_Q} \leq n_C$ and $n_D \leq n_{T^Q} + n_{D^Q}$. The first one can be deduced from Lemma 2, and the second one can be derived from Lemma 4.

- Let us now prove that $\overline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q) \geq \gamma(\mathbf{x}, \mathbf{y})$, for any $(\mathbf{x}, \mathbf{y}) \in \mathbb{E}(\mathbf{X}, \mathbf{Y})$. Let us set an arbitrary $(\mathbf{x}, \mathbf{y})$ and let C and D respectively denote the collection of concordant and discordant pairs in that sample. In order to prove the inequality $\overline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q) \geq \gamma(\mathbf{x}, \mathbf{y})$, let us first take into account that, according to Lemmas 2, 3 and 4, the following inequalities hold:

  - $n_C + n_T \leq n_{C^Q} + n_{T^Q}$.
  - $n_{D^Q} \leq n_D$.
  - $n_C + n_D + n_T = n_A = n_{C^Q} + n_{D^Q} + n_{T^Q}$.

According to this last equality, we can write:

$$\overline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q) = \frac{(n_{C^Q} + n_{T^Q}) - n_{D^Q}}{n_{C^Q} + n_{T^Q} + n_{D^Q}} = \frac{(n_{C^Q} + n_{T^Q}) - n_{D^Q}}{n_A}.$$

Furthermore, taking the first and the second inequalities into account, we get:

$$\frac{(n_C - n_D) + n_T}{(n_C + n_D) + n_T} = \frac{n_C + n_T - n_D}{n_A} \leq \frac{(n_{C^Q} + n_{T^Q}) - n_{D^Q}}{n_A}. \tag{6}$$

Now, from Lemma 5, we can deduce the following inequality:

$$\gamma(\mathbf{x}, \mathbf{y}) = \frac{(n_C - n_D)}{(n_C + n_D)} \leq \frac{(n_C - n_D) + n_T}{(n_C + n_D) + n_T}. \tag{7}$$

Thus, joining both inequalities (Equations 6 and 7), we get:

$$\gamma(\mathbf{x}, \mathbf{y}) = \frac{n_C - n_D}{n_C + n_D} \leq \frac{(n_{C^Q} + n_{T^Q}) - n_{D^Q}}{n_{C^Q} + n_{T^Q} + n_{D^Q}} = \overline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q). \quad \square$$

2.2.2. Optimality of bounds

Theorem 1 proves that the interval-valued Kendall's τ introduced in Definition 1 contains the set of all the feasible values for $\tau(\mathbf{x}, \mathbf{y}) \in \tau(\mathbb{E}(\mathbf{X}, \mathbf{Y}))$, where $(\mathbf{X}, \mathbf{Y})$ represents incomplete information about our original (non-quantized) sample $(\mathbf{x}, \mathbf{y})$.

But we can furthermore guarantee that the extremes of the interval-valued Kendal τ are included in the sets $\tau(\mathbb{E}(\mathbf{X}, \mathbf{Y}))$, $\gamma(\mathbb{E}(\mathbf{X}, \mathbf{Y}))$ and $\tau_b(\mathbb{E}(\mathbf{X}, \mathbf{Y}))$. We will now provide a constructive proof. In order to attain the upper bound $\overline{\tau}$, we will find a sample $(\mathbf{x}, \mathbf{y}) \in \mathbb{E}(\mathbf{X}, \mathbf{Y})$ wrt which all the pairs $(i, j) \in T^Q$ are concordant pairs. Analogously, the lower bound $\underline{\tau}$ will be reached by a sample that turns all pairs in T^Q into discordant pairs. This justifies the above formula where, in case of $\overline{\tau}$, the contributions of pairs in T^Q are counted positively, whereas in the case of $\underline{\tau}$ they are counted negatively. These two samples will not contain any tie, and therefore, the three variants of Kendall's coefficient, τ, τ_b and γ will take the same value for each of them.

We will make use of two auxiliary lemmas. The proof of the first one is immediate.

Lemma 6. *Let $i, j, k \in \{1, \ldots, n\}$ be three indices satisfying the restrictions $i < j$ and $k < j$.*

- *$(i, j) \in S_1^Q \cup U^Q$ if and only if ($X_i <_{ID} X_j$ or $X_i = X_j$), or equivalently iff $[\underline{x_i} \leq \underline{x_j}$ and $\overline{x_i} \leq \overline{x_j}]$.*
- *$(k, j) \in L_1^Q \cup U^Q$ if and only if ($X_j <_{ID} X_k$ or $X_j = X_k$), or equivalently iff $[\underline{x_j} \leq \underline{x_k}$ and $\overline{x_j} \leq \overline{x_k}]$.*

Lemma 7. *Consider two open intervals $]a,b[$ and $]c,d[$ satisfying the additional restrictions $a<d$ and $b>c$. Then,*

$$]c,d[\,\cap\,]a,b[\,\neq\emptyset.$$

Proof: According to the above hypothesis, both a and c are strictly lower than $\min\{b,d\}$, and therefore $\max\{a,c\}$ is strictly lower than $\min\{b,d\}$. Therefore we have $x\in\mathbb{R}$ satisfying:

$$\max\{a,c\}<x<\min\{b,d\},$$

which implies that $x\in]a,b[$ and $x\in]c,d[$.

Theorem 2. *$\underline{\tau}(\mathbf{x}^Q,\mathbf{y}^Q)$ and $\overline{\tau}(\mathbf{x}^Q,\mathbf{y}^Q)$ are respectively the minimum and maximum values of the three sets $\tau(\mathbb{E}(\mathbf{X},\mathbf{Y}))$, $\tau_b(\mathbb{E}(\mathbf{X},\mathbf{Y}))$ and $\gamma(\mathbb{E}(\mathbf{X},\mathbf{Y}))$.*

Proof: Let us describe how to reach the upper bound $\overline{\tau}(\mathbf{x}^Q,\mathbf{y}^Q)$ (A similar proof would apply to the lower bound). To do so, we will define a recursive selection procedure. Considering the two sequences of intervals, this selection procedure will consist of recursively picking, for each $j=1,\dots,n$, a pair of real numbers in $]\underline{x_j},\overline{x_j}[\times]\underline{y_j},\overline{y_j}[$ such that in the end all the pairs in T^Q correspond to pairs of vectors $((x^*_i,y^*_i),(x^*_j,y^*_j))$ which are concordant. Therefore, the three Kendall's coefficients τ,τ_b and γ will take the same value for this sample, and furthermore, the three of them will coincide with $\overline{\tau}(\mathbf{x}^Q,\mathbf{y}^Q)$.

Let us select an arbitrary pair of values $(x^*_1,y^*_1)\in]\underline{x_1},\overline{x_1}[\times]\underline{y_1},\overline{y_1}[$. Then, for all $j\in\{2,\dots,n\}$, we pick a new scalar pair $(x^*_j,y^*_j)\in]\underline{x_j},\overline{x_j}[\times]\underline{y_j},\overline{y_j}[$ satisfying the following conditions with respect to any previously selected (x^*_i,y^*_i) for $i<j$:

$$\max\{x^*_i:\ (i,j)\in S^Q_1\cup[U^Q\cap(L^Q_2)^c]\}<x^*_j<\min\{x^*_k:(k,j)\in L^Q_1\cup(U^Q\cap L^Q_2)\} \tag{8}$$

$$\max\{y^*_i:\ (i,j)\in S^Q_2\cup[V^Q\cap(L^Q_1)^c]\}<y^*_j<\min\{y^*_k:(k,j)\in L^Q_2\cup(V^Q\cap L^Q_1)\}, \tag{9}$$

where S^Q_i,L^Q_i,T^Q_i, $i=1,2$ correspond to the notation introduced in Lemma 1.

We will recursively prove that such a sequence of pairs (x^*_j,y^*_j), $j\in\{1,\dots,n\}$ can be constructed. In order to prove it, let us set an arbitrary $j\in\{2,\dots,n\}$. Let us assume that, for every $j'<j$, $x^*_{j'}$ belongs to the open interval $]\underline{x_{j'}},\overline{x_{j'}}[$ and satisfies

$$\max\{x^*_i:\ (i,j')\in S^Q_1\cup[U^Q\cap(V^Q\cup S^Q_2)]\}<x^*_{j'}<\min\{x^*_k:(k,j')\in L^Q_1\cup(U^Q\cap L^Q_2)\}, \tag{10}$$

and let us prove that following statements hold:

(a) $x^*_i<\overline{x_j}$ for every $(i,j)\in S^Q_1\cup[U^Q\cap(L^Q_2)^c]$.
Proof of a.- In fact, this set of pairs is included in $S^Q_1\cup U^Q$, and therefore, according to Lemma 6, $\overline{x_i}\leq\overline{x_j}$. Furthermore, $x^*_i<\overline{x_i}$ for every $i<j$ by assumption. Therefore, we derive the inequality $x^*_i<\overline{x_j}$.

(b) $x_k^* > \underline{x_j}$ for every k such that $(k,j) \in L_1^Q \cup (U^Q \cap L_2^Q)$.
Proof of b.- The proof is very similar to that of Part (a): First, let us note that the set of pairs $L_1^Q \cup (U^Q \cap L_2^Q)$ is included in $L_1^Q \cup U^Q$. Thus, according to Lemma 6, $\underline{x_k} \geq \underline{x_j}$. Furthermore, $x_k^* > \underline{x_k}$ by assumption, and therefore we get $x_k^* > \underline{x_j}$.

(c) $x_i^* < x_k^*$ for every i and k such that $(i,j) \in S_1^Q \cup [U^Q \cap (L_2^Q)^c]$, and $(k,j) \in L_1^Q \cup (U^Q \cap L_2^Q)$.
Proof of c.- Note first that, under the above restrictions, we have $(i,j) \in S_1^Q \cup U^Q$ and $(k,j) \in L_1^Q \cup U^Q$. Therefore, according to Lemma 6, the following inequalities hold:

$$\underline{x_i} \leq \underline{x_j} \leq \underline{x_k} \text{ and } \overline{x_i} \leq \overline{x_j} \leq \overline{x_k}.$$

We will divide the rest of the proof of (c) into three cases:

- Case 1: The case where $(i,j) \in S_1^Q$. Under this case, $\overline{x_i} \leq \underline{x_j}$, and therefore we get $x_i^* < \overline{x_i} \leq \underline{x_j} \leq \underline{x_k} < x_k^*$.
- Case 2: The case where $(k,j) \in L_1^Q$. Under this case, $\overline{x_j} \leq \underline{x_k}$, and therefore we get $x_i^* < \overline{x_i} \leq \overline{x_j} \leq \underline{x_k} < x_k^*$.
- Case 3: The case where $(i,j) \in (S_1^Q)^c$ and $(k,j) \in (L_1^Q)^c$. According to the above constraints, the only possibility for this case is that $(i,j) \in U^Q \cap (L_2^Q)^c$ and $(k,j) \in U^Q \cap L_2^Q$.
  - The condition $(i,j) \in U^Q$ and $(k,j) \in U^Q$ implies that $X_i = X_j = X_k$ and therefore either $(i,k) \in U^Q$ or $(k,i) \in U^Q$ (depending on whether $(i,k) \in A$ or $(k,i) \in A$).
  - Otherwise, the fact that $(i,j) \in (L_2^Q)^c$ and $(k,j) \in L_2^Q$ implies that $\overline{y_i} \leq \overline{y_j}$ and $\overline{y_j} \leq \underline{y_k}$, and therefore $\overline{y_i} \leq \underline{y_k}$. Therefore, either $i < k$ and therefore $(i,k) \in S_2^Q$ or $k < i$ and therefore $(k,i) \in L_2^Q$. In any of those cases, we can deduce that $x_i^* < x_k^*$ from Equation (10).

We can deduce from (a)-(c) that $l_j < \overline{x_j}$, $u_j > \underline{x_j}$ and $l_j < u_j$, where l_j and u_j respectively denote

$$l_j = \max\{x_i^* : (i,j) \in L_1^Q \cup (U^Q \cap L_2^Q)\}$$

and

$$u_j = \min\{x_i^* : (i,j) \in L_1^Q \cup (U^Q \cap L_2^Q)\}.$$

Therefore, according to Lemma 7, the intersection $(\underline{x_j}, \overline{x_j}) \cap (l_j, u_j)$ is non-empty. Thus, we have $x_j^* \in]\underline{x_j}, \overline{x_j}[$ satisfying Equation (8), i.e. satisfying $l_j < x_j^* < u_j$.

A similar constructive proof of that proposed here, in order to check the existence of $(x_1^*, \ldots, x_n^*)$, would be used to prove the existence of another sequence $(y_1^*, \ldots, y_n^*) \in Y_1 \times \ldots \times Y_n$ satisfying Equation (9) for every $j \in \{2, \ldots, n\}$.

We have proved that there exists a sample (x^*, y^*) with $x^* = (x_1^*, \ldots, x_n^*) \in X_1 \times \ldots \times X_n$ and $y^* = (y_1^*, \ldots, y_n^*) \in Y_1 \times \ldots \times Y_n$ respectively satisfying Equations 8 and 9, for every $j = 1, \ldots, n$. By construction, all the pairs $(i, j) \in T^Q$ are concordant in the constructed sample, and therefore any of the three Kendall coefficients associated to the sample coincides with $\overline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q)$. $\square$

Remark 2. *Note that our proof does not need any assumption on the linearity of the quantization process – i.e. the Φ_k^i $(i = 1, 2)$ can have different spread. The Φ_k^i simply must not overlap, which is a necessary property for a quantization process to lead to coherent digital images (signals).*

2.2.3. Link with the general interval-valued rank correlation coefficient

In the above results, we assumed that our incomplete information was derived from a quantization process, i.e. the interval valued sequences $\mathbf{X}$ and $\mathbf{Y}$ are induced by the quantization of unknown real valued sequences $\mathbf{x}$ and $\mathbf{y}$. We can easily adapt the respective definitions of $\underline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q)$ and $\overline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q)$ in order to get a similar result for the general case where the components of $(\mathbf{X}, \mathbf{Y})$ are not necessarily disjoint intervals.

Let $\underline{\tau}(\mathbf{X}, \mathbf{Y})$ and $\overline{\tau}(\mathbf{X}, \mathbf{Y})$ respectively denote:

$$\underline{\tau}(\mathbf{X}, \mathbf{Y}) = \frac{n_{C^Q} - n_{D^Q} - n_{T'^Q}}{n_A},$$

$$\overline{\tau}(\mathbf{X}, \mathbf{Y}) = \frac{n_{C^Q} - n_{D^Q} + n_{T'^Q}}{n_A}.$$

where C^Q and D^Q are in accordance with Equations 2 and 3 and T'^Q is defined as follows:

- $U'^Q = \{i, j) \in A \,|\, X_i \not<_{ID} X_j$ and $X_i \not>_{ID} X_j\}$
- $V'^Q = \{i, j) \in A \,|\, Y_i \not<_{ID} Y_j$ and $Y_i \not>_{ID} Y_j\}$
- $T'^Q = U'^Q \cup V'^Q$.

Let us first note that $\underline{\tau}(\mathbf{X}, \mathbf{Y})$ and $\overline{\tau}(\mathbf{X}, \mathbf{Y})$ respectively coincide with $\underline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q)$ and $\overline{\tau}(\mathbf{x}^Q, \mathbf{y}^Q)$ in the quantization case, since T'^Q coincides with T^Q in that particular setting. Thus, the above formulas just generalize Definition 1 by means of replacing the cardinal of

$$T^Q = \{(i, j) \in A \,|\, X_i = X_j \text{ or } Y_i = Y_j\}$$

by the cardinal of

$$T'^Q = \{(i, j) \in A \,|\, (X_i \not<_{ID} X_j \text{ and } X_i \not>_{ID} X_j) \text{ or } (Y_i \not<_{ID} Y_j \text{ and } Y_i \not>_{ID} Y_j)\}.$$

Let us now prove that $\underline{\tau}(\mathbf{X}, \mathbf{Y}) \leq \tau^-(\mathbf{X}, \mathbf{Y}) \leq \tau^+(\mathbf{X}, \mathbf{Y}) \leq \overline{\tau}(\mathbf{X}, \mathbf{Y})$: We can easily check that the two collections of sets $\{S_1^Q, L_1^Q, U'^Q\}$ and $\{S_2^Q, L_2^Q, V'^Q\}$ also form partitions of A and therefore the collection $\{C^Q, D^Q, T'^Q\}$ derived from them according to Equation 5, also forms a partition of A.

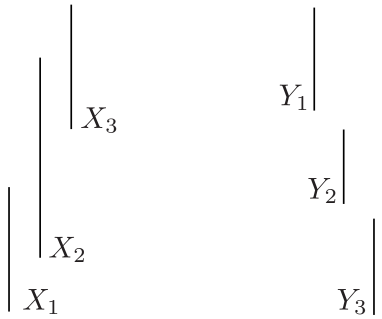

Figure 1: Example of a configuration where the bounds of $\overline{\underline{\tau}}$ are not reached.

Furthermore, given an arbitrary sample $(\mathbf{x}, \mathbf{y}) \in \mathbb{E}(\mathbf{X}, \mathbf{Y})$, and letting C, D and T respectively denote the sets of concordant, discordant and tied pairs in it, we can easily check that the following relations hold:

- $C^Q \subseteq C$
- $D^Q \subseteq D$,

and therefore, $C^Q \cup T'^Q \supseteq C \cup T$ and $D^Q \cup T'^Q \supseteq D \cup T$. A similar proof of that from Theorem 1 leads us to the fact that the interval $\overline{\underline{\tau}} = [\underline{\tau}(\mathbf{X}, \mathbf{Y}), \overline{\tau}(\mathbf{X}, \mathbf{Y})]$ calculated in this general setting also contains the three sets $\tau(\mathbb{E}(\mathbf{X}, \mathbf{Y}))$, $\tau_b(\mathbb{E}(\mathbf{X}, \mathbf{Y}))$ and $\gamma(\mathbb{E}(\mathbf{X}, \mathbf{Y}))$ in this general framework. Furthermore, these extreme points respectively coincide with $\underline{\tau}(\mathbf{x}^Q, \mathbf{x}^Q)$, and $\overline{\tau}(\mathbf{x}^Q, \mathbf{x}^Q)$ in the quantization case.

The fact that, in the general case, the bounds are not optimal, i.e. $\tau^-(\mathbf{X}, \mathbf{Y}) \neq \underline{\tau}(\mathbf{X}, \mathbf{Y})$ and $\tau^+(\mathbf{X}, \mathbf{Y}) \neq \overline{\tau}(\mathbf{X}, \mathbf{Y})$, is illustrated in the following example.

Example 1. *As we have already discussed, the maximization (resp. minimization) procedure relies on the fact that every pair in T^Q could be set as concordant (resp. discordant) independently of the others. The basis for this procedure relies on the following property:*

$$(i, j) \in T^Q, (j, k) \in T^Q \Rightarrow (i, k) \in T^Q.$$

But this does not necessarily happen when the observed sequences of intervals contain pairs of non-disjoint intervals.

Consider, for instance, the situation illustrated in Figure 1. According to the notation in Lemma 1 and according to the above definition of U'^Q, V'^Q and T'^Q we have:

- $S_1^Q = \{(1, 3)\}$,
- $L_1^Q = \emptyset$,
- $U'^Q = \{(1, 2), (2, 3)\}$, *and*

- $S_2^Q = \emptyset$
- $L_2^Q = \{(1,2),(1,3),(2,3)\}$,
- $V'^Q = \emptyset$,

and therefore:

- $C^Q = \emptyset$
- $D^Q = \{(1,3)\}$, *and*
- $T'^Q = \{(1,2),(2,3)\}$.

Thus, our interval-valued coefficient is $\underline{\overline{\tau}} = [-1, \frac{1}{3}]$. *However, the bound* $\frac{1}{3}$ *is unreachable, since it is impossible to get a sample where both pairs* $(1,2)$ *and* $(2,3)$ *are concordant (i.e., belong to* C*). This would require the pair* $(1,3)$ *to be concordant as well, something that is impossible, since* $(1,3) \in D^Q$ *(the collection of "necessarily" discordant pairs).*

3. Illustrative experiment

The following illustrative experiment aims to show that the rank correlation is not robust to quantization, and that using the imprecise τ we proposed is a rigorous alternative. In this experiment, we consider the two highly correlated images depicted in Figures (2.a) and (2.b). Those images were obtained by filling a Hoffman 2-D brain phantom (Data Spectrum Corporation) with a 99m technetium solution (148 MBq/L) and placing it on one of the detectors of a dual-head gamma camera using a low-energy high-resolution parallel-hole collimator (INFINIA, General Electric Healthcare). We obtained those two images by carrying out 2 acquisitions (acquisition time: 700 seconds; average count per image: 1000 kcounts). The values were quantized on 12 bits to produce the two images of Figures (2.a) and (2.b). As the acquisition time was quite long, the signal-to-noise ratio in the images was quite high.

The two images were highly correlated. As the brain phantom was symmetrical, the right and left part of the brain were also correlated, but slightly less due to the acquisition conditions. In order to lower the correlation coefficient (otherwise the upper bound of the imprecise τ would always reach one), we flipped the second image along the vertical axis (Figures (2.c)). Figure (3) shows the changes in the images due to quantization. The experiment involved varying the quantization level and each time comparing variations in the bounds of the imprecise τ, which are the possible "true" values of the correlation coefficient, and variations of τ, τ_b and γ. Figure (4) plots the results of this experiment.

We can see that τ, τ_b and γ yield incoherent results. They exhibit divergent behavior at low quantization levels, while being computed from quantized images corresponding to the same underlying real-valued image. Let us suppose that three experts have to decide whether or not the two images are correlated,

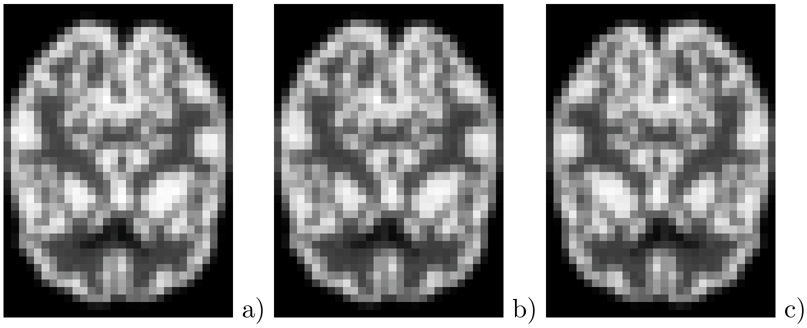

Figure 2: Two high dynamic images of the Hoffman 2-D brain phantom (a) and (b), image (b) after being vertically flipped (c).

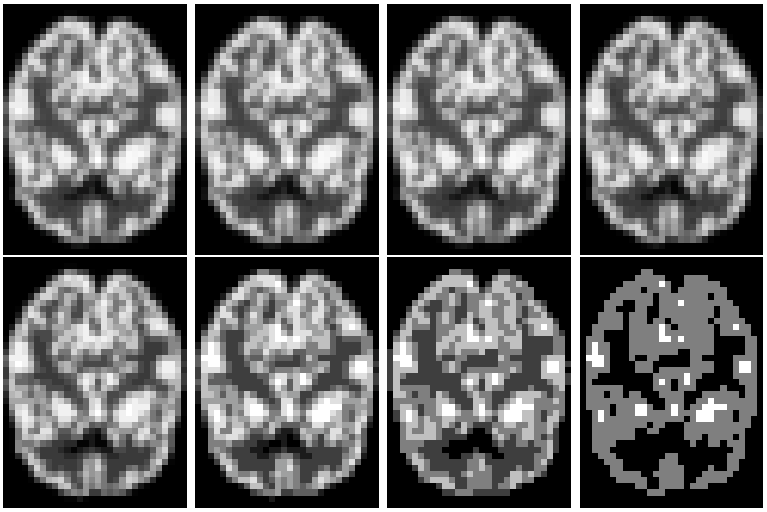

Figure 3: Different quantizations of the image depicted in Figure (2.a).

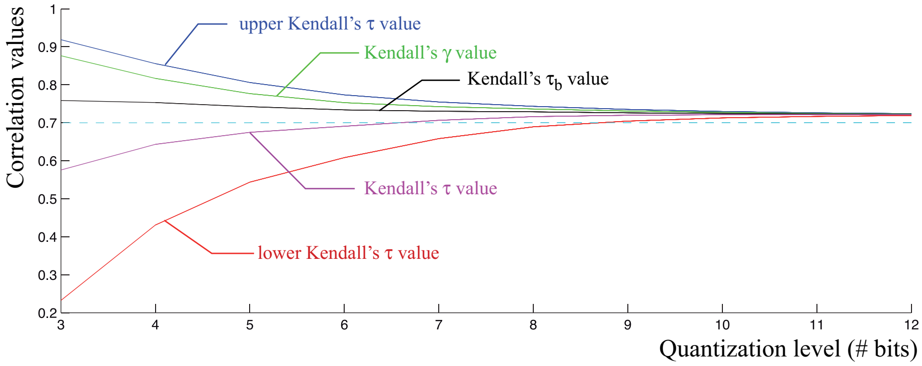

Figure 4: Variations in Kendall's τ with respect to the number of bits used to quantize the images.

with the first one using γ, the second one τ_b and the last one τ. Let us also suppose that, based on a set of experiments, they both agree that 0.8 can be considered as a good threshold for making this decision. Then, at a quantization level of 5, the first one will see a correlation while the second will not. However, the first and second will both both agree on no correlation when the quantization level is higher than 7. The last expert will never see a correlation. The decision of the τ_b expert will change according to the quantization level, although it is based on the same underlying image. This experiment reflects the arbitrariness of a correlation analysis based on quantized data. This arbitrariness is caused by the fact that quantized data is imprecise and thus a whole set of correlation values are possible when considering the real-valued digital image that underlies the quantized valued digital image.

On the other hand, using the above proposed $\overline{\underline{\tau}}$ enables us to consider all possible values of the correlation coefficient computed on the underlying real valued data. Using this imprecise correlation coefficient is easy. Let us consider that our expert is faced with the same problem of deciding whether the two underlying real-valued images are correlated or not based on the quantized images at different quantization levels. Above a quantization level of 7 he/she can conclude surely that the sequences are not correlated because the higher bound of $\overline{\underline{\tau}}$ is lower than the threshold (given that this threshold is relevant and that there are no uncertainties other than the data quantization). In that case, the expert's decision matches that of all three previous experts. But when the quantization level is lower, then a careful analyst would simply conclude that, given the state of the data, no meaningful conclusions can be drawn because, among the real valued images corresponding to the actual quantized image, some images are correlated and others are not. Adopting an imprecise point of view therefore yields much more robust and informed analysis and decisions. It advocates in favour of the possibility of non-decision, although not always yielding an analysis which cannot give conclusions. Non-decisions are in this context presented as a result of a rational analysis and they shall therefore not be considered as a failure of the expert to correctly interpret the data, but rather as a reflection of the irreducible uncertainty of the real world.

4. Discussion

The interval-valued $\overline{\underline{\tau}}(\mathbf{x}^Q, \mathbf{y}^Q)$ we propose aims at providing the exact bounds of all possible values of any variants of the Kendall's correlation coefficient of the set of all feasible pairs $(\mathbf{x}, \mathbf{y})$ of samples whose quantized values is the sequence $(\mathbf{x}^Q, \mathbf{y}^Q)$. Using this kind of coefficient is particularly relevant in image processing where data quantization could have a drastic influence in the final decision. In fact, comparing two digital images (signals) generally involves comparing the two original real-valued images (signals). Choosing to compute this correlation via τ^Q, τ_b^Q or γ^Q amounts to making a particular choice in the set of all feasible pairs defined in Subsection 2.2. Such a choice can lead to an arbitrary decision based on information that is not conveyed by the data, as illustrated in Section 3. Our approach is more cautious. As in [13], we propose to use this imprecise valued coefficient to make careful decisions. A decision based on $\overline{\underline{\tau}}(\mathbf{x}^Q, \mathbf{y}^Q)$ induces, as usual, a correlation threshold α. Comparing $\overline{\underline{\tau}}(\mathbf{x}^Q, \mathbf{y}^Q)$ to α can lead to an undecidability situation if $\alpha \in \overline{\underline{\tau}}(\mathbf{x}^Q, \mathbf{y}^Q)$ because there is at least one sequence in the feasible pairs providing a correlation coefficient value that is above α and one sequence in the feasible pairs providing a correlation coefficient value that is below α.

One can object that the approach proposed by [1] and [14] could be a good option since it leads to obtaining a better estimation of the correlation coefficient of two sequences of observations based on their noisy observation. Within this view of the problem, quantized sequences can be considered as noisy observations of the original sequences, and thus the fuzzy gamma rank correlation coefficient $\tilde{\gamma}$ they propose could potentially be a better choice for estimating the rank correlation coefficient than using τ^Q, τ_b^Q, γ^Q. We will now show that this approach to the problem is not relevant. Let $(\mathbf{x}, \mathbf{y})$ be the original sequences, $(\mathbf{x}^Q, \mathbf{y}^Q)$ their quantized observation and $(\mathbf{X}, \mathbf{Y})$ the interval valued sequence induced by the quantization process. Let $\mathbf{m}_x$ (rsp. $\mathbf{m}_y$) be the sequence of the mid-points of $\mathbf{X}$ (rsp. $\mathbf{Y}$), then it is more than likely (and generally recognized) that the sequences $\epsilon_x = (\mathbf{x} - \mathbf{m}_x)$ and $\epsilon_y = (\mathbf{y} - \mathbf{m}_y)$ are uniformly distributed. In fact, this approach could be suitable if $(\mathbf{x}, \mathbf{y})$ were considered as noisy observations of $(\mathbf{x}^Q, \mathbf{y}^Q)$, since the latter is more likely to be tied than the former. Since the observed values are quantized values, this random view of quantization is not relevant because the quantization is a fully deterministic observation process.

Moreover, when using $\tilde{\gamma}$ to estimate the correlation coefficient can lead to a biased decision. So let us consider the following example: $\mathbf{x}^Q = (1, 2, 5)$ and $\mathbf{y}^Q = (1, 5, 3)$. In that case, whatever the original feasible pairs $(\mathbf{x}, \mathbf{y})$, we have $n_C = 2$, $n_D = 1$, $n_T = 0$ and $n_A = 3$, leading to $\underline{\tau} = \tau = \tau_b = \gamma = \overline{\tau} = \frac{1}{3}$. Now, let us consider the $\top$-E-ordering based on the fuzzy relation $E(x, y) = \max(0, 1 - \frac{1}{r}|x - y|)$ and the Łukasiewicz t-norm (see [14] page 3) to compute $\tilde{\gamma}$ based on the quantized observations. In that case, if $r \leq 1$, then $\tilde{\gamma} = \overline{\tau} = \underline{\tau} = \frac{1}{3}$. However, if $r \in]1, 2]$, then $\tilde{n}_C = 2 - \delta$ (with $\delta = \frac{1}{r}$) and thus $\tilde{\gamma} = \frac{1-\delta}{3-\delta} < \frac{1}{3}$. Thus $\tilde{\gamma}$ does not correspond to any feasible pair of samples.

The situation considered in this article is a particular from the general case proposed by Denœux et al. ([3]) focused on situations where data imprecision is due to quantization. In this more specific case, the partial order over a set of intervals, considered by Denœux et al., is replaced by a strict total order over the set of intervals associated with the set of quantized levels (see Section 2.1). This approach leads to a simpler and more tractable problem that entails an easy estimation of the interval-valued rank correlation coefficient at low computational cost. Considering that, in this general case, the interval-valued coefficient $[\underline{\tau}, \overline{\tau}]$ we propose always bounds the set of feasible values for τ, this rough approximation can be used to estimate upper and lower bounds for it. In fact, the Monte-Carlo based estimation proposed in [6] and the heuristic proposed in [7] provide an inner estimation $[\tau_*, \tau^*]$ of $[\tau^-, \tau^+]$ (i.e. $\tau^- \leq \tau_* \leq \tau^* \leq \tau^+]$). Thus, both approaches can be used to compute an imprecise estimate $[\underline{\tau}, \tau_*]$ of τ^- and an imprecise estimate $[\tau^*, \overline{\tau}]$ of τ^+.

5. Conclusion

Here, we have proposed an interval-valued extension of Kendall's rank correlation, also called Kendall's τ. The use of intervals is prompted by the epistemic approach we take on the imprecision conveyed by the use of quantized values. This modeling assumption allows us to consider the set of all possible values before quantization without being forced to make any assumptions on the data, the linearity of the quantization process or on the existence of a probability distribution. The exactness of these bounds has been proven. An experimental illustration based on real-life data was proposed, which highlighted the need of considering the inherent imprecision of quantized data.

The more general case of rank correlation on imprecise data is more problematic and the bounds of our $\overline{\underline{\tau}}$ are only conservative due to the complex ranking configurations that result from partial overlapping of intervals. No authors have provided analytical ways of computing those bounds, and probably the best way to do so is by developing efficient general-purpose optimization methods.

Acknowledgements

The authors would like to thank Dr. Denis Mariano Goulart (Nuclear Medicine Department of the Montpellier Lapeyronie Hospital) for providing the data that was used in the experimental section.

This work was partially supported by TIN2014-56967-R (Spanish Ministry of Science and Innovation), FC-15-GRUPIN14-073 (Regional Ministry of the Principality of Asturias) and the French Embassy in Spain via the Merimée exchange program.

References

[1] U. Bodenhofer and F. Klawonn. Robust rank correlation coefficients on the basis of fuzzy orderings : Initial steps. *Mathware & Soft Computing*, 15:5 – 20, 2008.

[2] I. Couso and D. Dubois. Statistical reasoning with set-valued information: Ontic vs. epistemic views. *International Journal of Approximate Reasoning*, 55(7):1502–1518, 2014.

[3] T. Denoeux, M-H. Masson, and P-A. Hébert. Nonparametric rank-based statistics and significance tests for fuzzy data. *Fuzzy Sets and Systems*, 153(1):1 – 28, 2005.

[4] Y. Eugene and R. Johnston. The ineffectiveness of the correlation coefficient for image comparisons. *Technical Report LA-UR-96-2474, Los Alamos*, 1996.

[5] P. Fishburn. *Interval Orderings.* Wiley, 1987.

[6] P-A. Hébert, T. Denoeux, and M-H. Masson. Fuzzy rank correlation between fuzzy numbers. In *IFSA World Congress*, pages 224–227, 2003.

[7] O. Hryniewicz and K. Opara. Efficient calculation of kendall τ for interval data. In *Synergies of Soft Computing and Statistics for Intelligent Data Analysis, Advances in Intelligent Systems and Computing*, volume 190, pages 203–210. Springer, 2013.

[8] O. Hryniewicz and A. Szediw. Fuzzy kendall τ statistic for autocorrelated data. In *Soft Methods for Handling Variability and Imprecision*, volume 44, pages 155–162. Springer, Berlin, 2008.

[9] G. Kordelas and P. Daras. Robust sift-based feature matching using kendall's rank correlation measure. In *Proceedings of the International Conference on Image Processing*, pages 325–328, 2009.

[10] J-P. Lecoutre and P. Tassi. *Statistique non-paramétrique et robustesse.* Economie et statistiques avancées. Economica, 1987.

[11] A. Mittal and V. Ramesh. An intensity-augmented ordinal measure for visual correspondence. In *IEEE Computer Conference on Computer Vision and Pattern Recognition*, volume 1, pages 849–856, 2006.

[12] K. Pearson. Contributions to the mathematical theory of evolution. *Proceedings of the Royal Society of London, Philosophical Transactions of the Royal Society*, 185:71110, 1984.

[13] J. Perolat, I. Couso, K. Loquin, and O. Strauss. Generalizing the wilcoxon rank-sum test for interval data. *International Journal of Approximate Reasoning*, 56:108 – 121, 2015.

[14] M-D. Ruiz and E. Hüllermeier. A formal and empirical analysis of the fuzzy gamma rank correlation coefficient. *Information Sciences*, 206:1–17, 2012.

[15] G. P. Sillito. The distribution of kendall's τ coefficient of rank correlation in rankings containing ties. *Biometrika*, 34:36–40, 1947.