A Formal and Empirical Analysis of the Fuzzy Gamma Rank Correlation Coefficient

Maria Dolores Ruiz[a], Eyke Hüllermeier[b]

[a] *Department of Computer Science and Artificial Intelligence*
University of Granada, Spain
[b] *Department of Mathematics and Computer Science*
Marburg University, Germany

Abstract

The so-called gamma coefficient is a well-known rank correlation measure frequently used to quantify the strength of dependence between two variables with ordered domains. To increase the robustness of this measure toward noise in the data, a generalization of the gamma coefficient has recently been developed on the basis of fuzzy order relations. The goal of this paper is threefold. First, we analyze some formal properties of the fuzzy gamma coefficient. Second, we complement the original experiments, which have been conducted on a simple artificial data set, by a more extensive empirical evaluation using real-world data. On the basis of these empirical results, we provide some basic insights and offer an explanation for the effectiveness of the fuzzy gamma coefficient. Third, we propose an alternative motivation for the measure, based on the idea of (fuzzy) equivalence relations induced by limited precision in the perception of measurements.

Keywords: rank correlation, gamma rank correlation measure, fuzzy order relations, fuzzy rank correlation, noisy data, equivalence relations

1. Introduction

Rank correlation measures have been studied extensively in non-parametric statistics and are routinely used in diverse fields of application, including clustering [3], information retrieval [15, 18], data mining [9, 16], economics [1, 17], and image processing [4, 2]. In contrast to numerical correlation measures such as the Pearson correlation, rank correlation measures are only based on the ordering of the observed values of a variable. Thus, measures of this kind are more widely applicable, not only to numerical but also to non-numerical variables with an ordered domain (i.e., measured on an ordinal scale).

Roughly speaking, when deriving a rank correlation, each value is first replaced by its rank, and the correlation is then computed on the rank data thus obtained. Mapping numerical values to their ranks does of course produce a certain loss of information. In many cases, this is intended and makes perfect sense, for example if a numerical variable does not have a natural (or unique) scale of measurement[1] or may contain outliers that would strongly bias a numerical correlation measure. On the other hand, a very small difference between two values may no longer be distinguished from a very large difference, since both could be mapped to adjacent ranks and, therefore, have an equal (transformed) distance of 1. Moreover, the robustness toward outliers, i.e., extremely small or large values, comes at the price of a certain sensitivity toward small changes: In many cases, a small increase or decrease of a value will not have any effect at all, but in some cases it may induce

[1] Note, for example, that a non-numerical transformation (like a log-transformation) of one variable will change its correlation with another variable.

a swapping of adjacent ranks. Thus, a small change between the numerical values can be "boosted" to a distance of 1 between the associated rank values. Obviously, this property is especially undesirable in the presence of noise in the data.

To overcome problems of this kind, Bodenhofer and Klawonn [8] have recently proposed a fuzzy variant of a rank correlation measure known as Goodman and Kruskal's gamma measure [12]. Roughly speaking, the use of fuzzy order relations allows the authors to distinguish between negligible and significant differences between numerical values in a more subtle way, and to decrease the influence of the former. Thus, the rank correlation measure becomes arguably more robust toward noise.

The goal of this paper is threefold. First, we analyze some formal properties of the fuzzy rank correlation measure proposed by Bodenhofer and Klawonn. Second, we complement the authors' experiments, which have been conducted on a simple artificial data set, by a more extensive empirical evaluation using real-world data. On the basis of these empirical results, we provide some basic insights and offer an explanation for the effectiveness of the fuzzy gamma coefficient.

Third, we offer an alternative motivation of the measure, based on the idea of equivalence relations induced by limited precision in the perception of measurements. As an illustrating example, suppose we are interested in the correlation, if any, between the length of a submitted manuscript and the recommendation of the reviewer. Since the recommendation is taken from an ordinal scale (e.g., accept, minor revision, major revision, reject), only a rank correlation measure can be computed. As will be explained in more detail

later on, such measures are mainly based on the order relation between two measurements. In the case of the recommendation scale, this order relation is simply defined by the ordinal scale; for example, accept $>$ minor revision. As for the length of a manuscript, one may simply compare two papers in terms of the respective number of words. This approach, however, is unlikely to capture the reviewer's perception. For example, a reviewer will normally not perceive an article A as longer than an article B, only because the former has one or two words more than the latter. In this situation, a fuzzy order relation (on the word count of manuscripts) can be used in a quite reasonable way, namely for expressing that an article A can be longer than B "to some degree". In other words, it allows for modeling the "perceived difference in length" between two articles as a gradual relation, which is arguably more natural than treating it in a binary way.

The above example does also hint at another appealing property of fuzzy rank correlation, namely the fact that they combine properties of both, numerical and rank correlation. Thus, like in the example, it becomes possible to compare variables that are measured on scales of different types. In the remainder of the paper, we shall mainly focus on the case where both variables are numeric, mainly because this case was also studied by Bodenhofer and Klawonn. One should keep in mind, however, that the approach is in principle more general and only requires the possibility to equip a domain with a reasonable fuzzy equivalence relation (note that the canonical $>$ relation determined by an ordinal scale can be seen as a degenerate fuzzy relation).

The remainder of the paper is organized as follows. In the next section, we recall the essential background for understanding the rest of the paper,

including rank correlation measures, fuzzy order relations, and the fuzzy extension of the gamma coefficient originally introduced in [8]. In Section 3, we derive some results throwing light on formal properties of the fuzzy gamma. In Section 4, we elaborate on the idea of using the fuzzy gamma as a noise-tolerant version of the original gamma coefficient. An alternative interpretation of the fuzzy gamma in terms of a "perception-based" rank correlation measure masking inappropriate precision in the measurement of quantities is then proposed and investigated in Section 5. Finally, we conclude the paper with a couple of remarks and an outlook on future work in Section 6.

2. Rank Correlation Coefficients

In this section, we give a brief overview of rank correlation measures in general and then focus on the gamma coefficient. We start with some formal definitions which are important to understand the rest of the paper. We also address the use of rank correlation coefficients as distance measures.

2.1. Basic Correlation Measures

A rank correlation measure is applied to $n \geq 2$ paired observations

$$\{ (x_i, y_i) \}_{i=1}^{n} \subset (\mathbb{X} \times \mathbb{Y})^n \tag{1}$$

of a pair of variables (X, Y), where $\mathbb{X}$ and $\mathbb{Y}$ are two linearly ordered domains (e.g., subsets of the reals); we denote $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$. The goal is to measure the dependence between the two variables in terms of their tendency to increase and decrease in the same or the opposite direction. If an increase in X tends to come along with an increase

in Y, then the (rank) correlation is positive. The other way around, the correlation is negative if an increase in X tends to come along with a decrease in Y. If there is no dependency of either kind, the correlation is (close to) 0.

Among the best-known and most frequently used measures are Spearman's rank correlation coefficient (Spearman's rho for short), Kendall's tau and Goodman and Kruskal's gamma. *Spearman's rho* is given by the sum of squared rank distances, normalized to the range $[-1, 1]$:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} (r(x_i) - r(y_i))^2}{n(n^2 - 1)} \ , \tag{2}$$

where $r(x_i) = \#\{j \in \{1, \ldots, n\} \,|\, x_j \leq x_i\}$ is the rank of value x_i in the set of observations $\{x_1, \ldots, x_n\}$. Here, we assume that the data does not contain any ties, i.e., $x_i \neq x_j$ for $1 \leq i \neq j \leq n$. In the presence of ties, a proper generalization of (2) can be used.

Kendall's tau coefficient is defined in terms of the number of concordant, discordant, and tied data points. For a given index pair $(i, j) \in \{1, \ldots, n\}^2$, we say that (i, j) is *concordant* if $x_i < x_j$ and $y_i < y_j$ or $x_i > x_j$ and $y_i > y_j$; it is *discordant* if $x_i < x_j$ and $y_i > y_j$ or $x_i > x_j$ and $y_i < y_j$; it is a *tie* if either $x_i = x_j$ or $y_i = y_j$. Denoting

$$\begin{aligned} C &= \#\{(i,j) \mid i < j,\ x_i < x_j \text{ and } y_i < y_j \text{ or } x_i > x_j \text{ and } y_i > y_j\}, \\ D &= \#\{(i,j) \mid i < j,\ x_i < x_j \text{ and } y_i > y_j \text{ or } x_i > x_j \text{ and } y_i < y_j\}, \\ T &= \#\{(i,j) \mid i < j,\ x_i = x_j \text{ or } y_i = y_j\}, \end{aligned} \tag{3}$$

the original Kendall tau is defined as

$$\tau_a = \frac{C - D}{\frac{1}{2} n(n-1)} \ . \tag{4}$$

When there are no ties ($T = 0$) and the two rankings coincide, we have $\frac{1}{2} n(n-1)$ concordant pairs and no discordant pair, so $\tau_a = 1$; if one ranking is the

reverse of the other one, we have $\tau_a = -1$. In the presence of ties, however, this measure does not assume the extreme values -1 and $+1$ and, hence, is not well normalized. Like in the case of Spearman's rho, a generalization can be defined by properly adapting the normalizing constant in (4).

Another quite simple measure is Goodman and Kruskal's *gamma rank correlation* [12], which simply ignores all ties. It is defined as

$$\gamma = \frac{C - D}{C + D} \tag{5}$$

and coincides with Kendall's tau when there are no ties in the data. Throughout the remainder of the paper, we shall focus on the measure (5).

2.2. Fuzzy Equivalence and Order Relations

Bodenhofer and Klawonn [8] advocate the use of the gamma coefficient as a reasonable correlation measure but also indicate problems in the presence of noise in the data. To make it more robust toward noisy data, they propose a fuzzy generalization which is based on concepts of fuzzy orderings and $\top$-equivalence relations, where $\top$ denotes a triangular norm (t-norm) [6, 8]. We assume that the reader is familiar with the basic concepts of triangular norms and fuzzy relations [6]. Yet, to make the paper more self-contained, we briefly recall some basic definitions which are necessary for the understanding of the rest of the paper.

Definition 1. A fuzzy relation $E : \mathbb{X} \times \mathbb{X} \to [0, 1]$ is called *fuzzy equivalence* with respect to a t-norm $\top$, for brevity $\top$-*equivalence*, if and only if it satisfies the following three axioms: For all $x, y, z \in \mathbb{X}$,

(i) reflexivity: $E(x, x) = 1$,

(ii) symmetry: $E(x,y) = E(y,x)$,

(iii) $\top$-transitivity: $\top(E(x,y), E(y,z)) \leq E(x,z)$.

Definition 2. A fuzzy relation $L : \mathbb{X} \times \mathbb{X} \to [0,1]$ is called *fuzzy ordering* with respect to a t-norm $\top$ and a $\top$-equivalence $E : \mathbb{X} \times \mathbb{X} \to [0,1]$, for brevity $\top$-*E-ordering*, if and only if it satisfies the following three axioms: For all $x, y, z \in \mathbb{X}$,

(i) E-reflexivity: $E(x,y) \leq L(x,y)$,

(ii) $\top$-E-antisymmetry: $\top(L(x,y), L(y,x)) \leq E(x,y)$,

(iii) $\top$-transitivity: $\top(L(x,y), L(y,z)) \leq L(x,z)$.

Moreover, we call a $\top$-E-ordering L *strongly complete* if $\max(L(x,y), L(y,x)) = 1$ for all $x, y \in \mathbb{X}$. Considering the special cases of the well-known and frequently used Łukasiewicz t-norm $\top_L$ and the product t-norm $\top_P$, defined by

$$\begin{aligned} \top_L(x,y) &= \max(0, x+y-1), \\ \top_P(x,y) &= xy, \end{aligned}$$

it can be verified that

$$\begin{aligned} E_r(x,y) &= \max\left(0, 1 - \frac{1}{r}\,|x-y|\right) \\ E'_r(x,y) &= \exp\left(-\frac{1}{r}\,|x-y|\right) \end{aligned} \tag{6}$$

are $\top$-equivalences on $\mathbb{R}$ associated with $\top_L$ and $\top_P$, respectively, where $r > 0$. The following theorem proved in [5] gives a full characterization of strongly complete orderings.

Theorem 1. (Bodenhofer [5]). Let L be a binary fuzzy relation on $\mathbb{X}$ and let E be a $\top$-equivalence on $\mathbb{X}$. Then the following two statements are equivalent:

(i) L is a strongly complete $\top$-E-ordering on $\mathbb{X}$.

(ii) There exists a linear ordering $\preceq$ such that relation E is compatible[2] with $\preceq$ and, moreover, L can be represented as follows:

$$L(x,y) = \begin{cases} 1 & \text{if } x \preceq y \\ E(x,y) & \text{otherwise} \end{cases} \tag{7}$$

This theorem implies that

$$L_r(x,y) = \min\left\{1, \max\left(0, 1 - \frac{1}{r}(x-y)\right)\right\}$$

is a strongly complete $\top_L$-E_r-ordering on $\mathbb{R}$, and

$$L'_r(x,y) = \min\left\{1, \exp\left(-\frac{1}{r}(x-y)\right)\right\}$$

is a strongly complete $\top_P$-E'_r-ordering on $\mathbb{R}$.

Definition 3. A binary fuzzy relation R is called a *strict fuzzy ordering* with respect to a t-norm $\top$ and a $\top$-equivalence E, or *strict* $\top$-E-*ordering* for short, if R is irreflexive ($R(x,x) = 0$ for all $x \in X$), $\top$-transitive and E-extensional, which means that

$$\top(E(x,x'), E(y,y'), R(x,y)) \leq R(x',y')$$

for all $x, x', y, y' \in \mathbb{X}$ [7].

[2] A fuzzy relation E is compatible with an order relation $\preceq$ on X if and only if $E(x,z) \leq \min\{E(x,y), E(y,z)\}$ holds for all $x \preceq y \preceq z$.

As argued in [7], the most appropriate way of extracting a strict fuzzy ordering R from a $\top$-E-ordering L is to define

$$R(x,y) = \min\{L(x,y), N(L(y,x))\} \ , \tag{8}$$

where $N(x) = \sup\{y \in [0,1] \,|\, \top(x,y) = 0\}$ is the residual negation of $\top$. Examples of this construction are the relations

$$R_r(x,y) = \min\left\{1, \max\left\{0, \frac{1}{r}(y-x)\right\}\right\},$$
$$R'_r(x,y) = \max\left\{0, 1-\exp\left(-\frac{1}{r}(y-x)\right)\right\}.$$

For a strongly complete $\top_L$-E-ordering L, the relation (8) is given by $R(x,y) = 1-L(y,x)$; moreover, $R(x,y)+E(x,y)+R(y,x) = 1$ and $\min\{R(x,y), R(y,x)\} = 0$.

2.3. A Fuzzy Extension of the Gamma Rank Correlation

Consider a set of paired data points (1) and assume to be given two $\top_L$-equivalences $E_{\mathbb{X}} : \mathbb{X}^2 \to [0,1]$ and $E_{\mathbb{Y}} : \mathbb{Y}^2 \to [0,1]$, a strongly complete $\top_L$-$E_{\mathbb{X}}$-ordering $L_{\mathbb{X}} : \mathbb{X}^2 \to [0,1]$ and a strongly complete $\top_L$-$E_{\mathbb{Y}}$-ordering $L_{\mathbb{Y}} : \mathbb{Y}^2 \to [0,1]$. We can then define a strict $\top_L$-$E_{\mathbb{X}}$-ordering on $\mathbb{X}$ by $R_{\mathbb{X}}(x_1,x_2) = 1-L_{\mathbb{X}}(x_2,x_1)$ and a strict $\top_L$-$E_{\mathbb{Y}}$-ordering on $\mathbb{Y}$ by $R_{\mathbb{Y}}(y_1,y_2) = 1-L_{\mathbb{Y}}(y_2,y_1)$. Using these relations, the concepts of concordance and discordance of data points can be generalized as follows: Given an index pair (i,j), the degree to which this pair is concordant, discordant, and tied is defined, respectively, as

$$\tilde{C}(i,j) = \top(R_{\mathbb{X}}(x_i,x_j), R_{\mathbb{Y}}(y_i,y_j)), \tag{9}$$
$$\tilde{D}(i,j) = \top(R_{\mathbb{X}}(x_i,x_j), R_{\mathbb{Y}}(y_j,y_i)), \tag{10}$$
$$\tilde{T}(i,j) = \bot(E_{\mathbb{X}}(x_i,x_j), E_{\mathbb{Y}}(y_i,y_j)), \tag{11}$$

where $\top$ is a t-norm and $\perp$ is the dual t-conorm of $\top$ (i.e. $\perp(x,y) = 1 - \top(1-x, 1-y)$). The following equality holds for all index pairs (i,j):

$$\tilde{C}(i,j) + \tilde{C}(j,i) + \tilde{D}(i,j) + \tilde{D}(j,i) + \tilde{T}(i,j) = 1.$$

Adopting the simple sigma-count principle to measure the cardinality of a fuzzy set [10], the number of concordant and discordant pairs can be computed, respectively, as

$$\tilde{C} = \sum_{i=1}^{n} \sum_{j \neq i} \tilde{C}(i,j), \qquad \tilde{D} = \sum_{i=1}^{n} \sum_{j \neq i} \tilde{D}(i,j).$$

The *fuzzy ordering-based* gamma rank correlation measure $\tilde{\gamma}$, or simply fuzzy gamma, is then defined as

$$\tilde{\gamma} = \frac{\tilde{C} - \tilde{D}}{\tilde{C} + \tilde{D}} \ . \tag{12}$$

Note that this measure is "parameterized" by the underlying fuzzy orderings, i.e., a t-norm $\top$ and fuzzy equivalence relations $E_{\mathbb{X}}$ and $E_{\mathbb{Y}}$.

From the definition of $\tilde{\gamma}$, it is clear that the basic idea is to decrease the influence of "close-to-tie" pairs (x_i, y_i) and (x_j, y_j). Roughly speaking, such pairs, whether concordant or discordant, are turned into a partial tie, and hence are ignored to some extent. Or, stated differently, there is a smooth transition between being concordant (discordant) and being tied. The larger the scaling parameter r, the more a pair is considered as a partial tie; see Fig. 1 for an illustration of the difference between the crisp and the fuzzy case.

As a side remark, we note that a fuzzy equivalence relation E may have a probabilistic interpretation, although this is not required by the formal framework. Consider, for example, the case of numerical data corrupted with

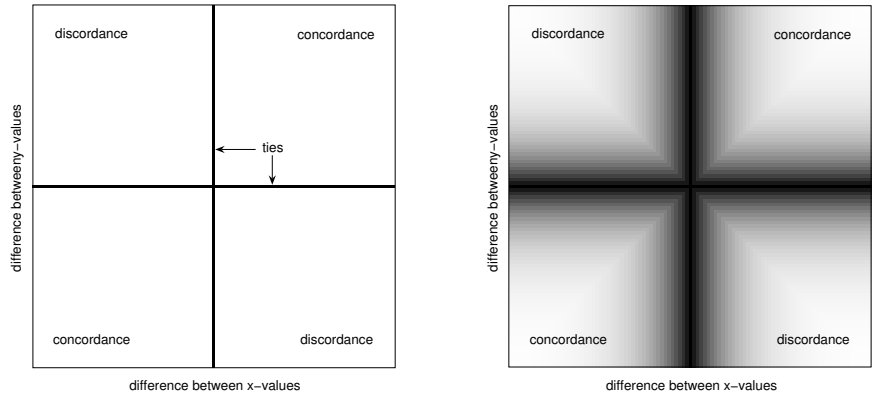

Figure 1: In the crisp (non-fuzzy) case, shown left, two data pairs (x_i, y_i) and (x_j, y_j) are either concordant or discordant, except for $x_i = x_j$ or $y_i = y_j$. Consequently, the set of ties coincides with two crossing lines in the left figure, where either the difference $x_i - x_j$ or the difference $y_i - y_j$ vanishes. In the fuzzy case, shown on the right, a data pair can be tied to a certain degree, as indicated by the level of gray.

an additive noise term ϵ: instead of the true value $\bar{x}$, only a noisy version $x = \bar{x} + \epsilon$ is observed. Moreover, suppose that the distribution D of ϵ is known. The degree of equivalence $E(x, x')$ of two observations $x \leq x'$ may then be defined as a (decreasing) function of the probability $\mathbf{P}(\bar{x} < \bar{x}') = \mathbf{P}(x - \epsilon < x' - \epsilon') = \mathbf{P}(\epsilon' - \epsilon < x' - x)$, where the latter corresponds to the value of the cumulative distribution function of the convolution $D - D$ at $x' - x$. Roughly speaking, the idea is that the more certain one can be that $\bar{x}$ is indeed smaller than $\bar{x}'$ (i.e., the larger the probability $\mathbf{P}(\bar{x} < \bar{x}')$), the less equivalent these values should be. For example, if ϵ is uniformly distributed in $[-c, +c]$, then $D - D$ has a triangular-shaped density $z \mapsto \max\left(\frac{1}{2c} - \frac{1}{4c^2} \cdot z, 0\right)$, and thus naturally suggests the modeling of E in terms of a triangular fuzzy set.

3. Formal Properties of Gamma and Fuzzy Gamma

3.1. Metric Properties

Consider a set $\mathbb{X}$ endowed with a total order; without loss of generality, we can assume that $\mathbb{X}$ is a subset of the reals. Ideally, a rank correlation measure C should satisfy the following for all $n \in \mathbb{N}$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{X}^n$:

C1: $-1 \leq C(\boldsymbol{x}, \boldsymbol{y}) \leq 1$

C2: $C(\boldsymbol{x}, \boldsymbol{y}) = C(\boldsymbol{y}, \boldsymbol{x})$

C3: $C(\boldsymbol{x}, \boldsymbol{y}) = 1$ if the elements in $\boldsymbol{x}$ are in the same order as those in $\boldsymbol{y}$, i.e., $r(\boldsymbol{x}) = r(\boldsymbol{y})$; in particular, $C(\boldsymbol{x}, \boldsymbol{x}) = 1$.

It is quite easy to see that γ satisfies all of these properties. Note that property C3 is fulfilled by all measures that only depend on the ranks $r(\boldsymbol{x}) = (r(x_1), r(x_2) \dots r(x_n))$ and $r(\boldsymbol{y}) = (r(y_1), r(y_2) \dots r(y_n))$. This property is intentionally violated by the fuzzy variant $\tilde{\gamma}$.

The above properties may remind one of related properties of distance measures, and indeed, some rank correlation measures are in fact normalized versions of corresponding distance measures. For example, Spearman's rho is an affine transformation of the sum of squared rank distances to the interval $[-1, +1]$, and Kendall's tau is a similar transformation of the Kendall distance, namely the sum of rank inversions [13]. To study the rank correlation measures γ and $\tilde{\gamma}$ from the perspective of a distance measure, recall the basic definition of a metric:

Definition 4. A mapping $d : A \times A \to \mathbb{R}$ is a metric on A if and only if it fulfills the following for all $a, b, c \in A$:

– $d(a, b) \geq 0$ (non-negativity),

– $d(a,b) = 0 \Leftrightarrow a = b$ (separation),

– $d(a,b) = d(b,a)$ (symmetry),

– $d(a,c) \leq d(a,b) + d(b,c)$ (triangle inequality).

Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{X}^n$. The measure

$$d(\boldsymbol{x}, \boldsymbol{y}) = 1 - \gamma(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{C - D}{C + D} = \frac{2D}{C + D}$$

is obviously non-negative (since $-1 \leq \gamma \leq 1$) and symmetric (since γ is also symmetric). The separation property cannot be satisfied, as a rank correlation measure depends on the concrete values $(x_i, y_i) \in \mathbb{X}^2$ only indirectly via the corresponding ranks. It holds, however, that $d(\boldsymbol{x}, \boldsymbol{y}) = 0$ implies $r(\boldsymbol{x}) = r(\boldsymbol{y})$.

It is also easy to see that the triangle inequality does not hold for d. Here is a simple counterexample: With $\boldsymbol{x} = (1, 2, 3)$, $\boldsymbol{y} = (1, 1, 2)$ and $\boldsymbol{z} = (2, 1, 2)$, we have $\gamma(\boldsymbol{x}, \boldsymbol{y}) = 1$, $\gamma(\boldsymbol{x}, \boldsymbol{z}) = 0$, $\gamma(\boldsymbol{y}, \boldsymbol{z}) = 1$, and hence $1 = d(\boldsymbol{x}, \boldsymbol{z}) \not\leq d(\boldsymbol{x}, \boldsymbol{z}) + d(\boldsymbol{z}, \boldsymbol{y}) = 0$. As a main reason for the violation of this inequality, note that, due to the ignorance of ties, the pairwise distance computations may refer to completely different elements. For example, since y_1 and y_2 are tied, $\gamma(\boldsymbol{x}, \boldsymbol{y})$ is completely determined by the comparison of the index pairs $(1, 3)$ and $(2, 3)$. Likewise, since z_1 and z_3 are tied, $\gamma(\boldsymbol{y}, \boldsymbol{z})$ only depends on the index pair $(2, 3)$, while $\gamma(\boldsymbol{x}, \boldsymbol{z})$ depends on the index pairs $(1, 2)$ and $(2, 3)$.

Replacing γ by the fuzzy version $\tilde{\gamma}$, we obtain

$$\tilde{d}(\boldsymbol{x}, \boldsymbol{y}) = 1 - \tilde{\gamma}(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{\tilde{C} - \tilde{D}}{\tilde{C} + \tilde{D}} = \frac{2\tilde{D}}{\tilde{C} + \tilde{D}} \ .$$

Again, it is obvious that $\tilde{d}(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ (because $-1 \leq \tilde{\gamma}(\boldsymbol{x}, \boldsymbol{y}) \leq 1$) and that symmetry holds (symmetry of a t-norm implies symmetry of $\tilde{C}$ and $\tilde{D}$).

Moreover, when comparing $\boldsymbol{x}$ with itself, the degree of discordance of the index pair (i, j) is

$$\begin{aligned}
\top(R_{\mathbb{X}}(x_i, x_j), R_{\mathbb{X}}(x_j, x_i)) &\leq R_{\mathbb{X}}(x_i, x_i) \\
&= 1 - L_{\mathbb{X}}(x_i, x_i) \\
&= 1 - \begin{cases} 1 & \text{if } x_i \leq x_i \\ E(x_i, x_i) & \text{otherwise} \end{cases} \\
&= 1 - 1 = 0
\end{aligned}$$

where we have used the $\top$-transitivity of $R_{\mathbb{X}}$, Theorem 1 and the reflexivity of E. (Note that, in particular, L can be written as $L(x, y) = \min(1, E(x, y))$, as can be seen for the cases of Łukasiewicz and product t-norms). Therefore, the total degree of discordance is 0, which means that reflexivity also holds.

Like in the non-fuzzy case, the triangle inequality does of course not hold. To show this, the same counterexample as above can be used. In fact, with $\boldsymbol{x} = (1, 2, 3)$, $\boldsymbol{y} = (1, 1, 2)$, $\boldsymbol{z} = (2, 1, 2)$, $\tilde{\gamma}(\boldsymbol{x}, \boldsymbol{y})$ is nearly zero while $\tilde{\gamma}(\boldsymbol{x}, \boldsymbol{z})$ and $\tilde{\gamma}(\boldsymbol{y}, \boldsymbol{z})$ are close to 1 if $0 < r < 1$, both for the Łukasiewicz and the product t-norm.

In summary, it can be seen that γ satisfies desirable properties of a rank correlation measure, though not all properties of a metric (especially not the triangle inequality). More importantly, however, none of the properties satisfied by γ are lost when passing to the fuzzy version $\tilde{\gamma}$ except, of course, property C3.

3.2. Limit Behavior of the Fuzzy Gamma

The concrete values produced by the $\tilde{\gamma}$ measure depend on the scaling parameter r. In the following, we study the influence of r on the relation between $\tilde{\gamma}$ and γ for the cases of the product and the Łukasiewicz t-norm. More specifically, we show that the natural requirement of recovering the original γ for the case $r = 0$ is indeed satisfied. More formally: $\tilde{\gamma}$ converges to γ as $r \to 0$.

Proposition 1. Let γ, $\tilde{\gamma}_L$ and $\tilde{\gamma}_P$ be defined as in the previous section. The following properties are satisfied:

i. $\lim_{r\to 0} \tilde{\gamma}_L = \gamma$
ii. $\lim_{r\to 0} \tilde{\gamma}_P = \gamma$

Proof. i. We first compute some simpler limits:

- If $x_j - x_i < 0$ then $\lim_{r\to 0} 1 - \frac{1}{r}(x_j - x_i) = \infty$, so $\lim_{r\to 0} L_{\mathbb{X}}(x_j, x_i) = 1$.
- If $x_j - x_i > 0$ then $\lim_{r\to 0} 1 - \frac{1}{r}(x_j - x_i) = -\infty$, so $\lim_{r\to 0} L_{\mathbb{X}}(x_j, x_i) = 0$.
- If $x_j - x_i = 0$ then we have a tie.

We distinguish four cases when taking the limit $r \to 0$:

(a) If $x_j - x_i < 0$ and $y_j - y_i < 0$, then $\lim_{r\to 0} L_{\mathbb{X}}(x_j, x_i) = 1$ and $\lim_{r\to 0} L_{\mathbb{Y}}(y_j, y_i) = 1$, therefore $\lim_{r\to 0} \tilde{C}(i, j) = 0$.

(b) If $x_j - x_i < 0$ and $y_j - y_i > 0$, then $\lim_{r\to 0} L_{\mathbb{X}}(x_j, x_i) = 1$ and $\lim_{r\to 0} L_{\mathbb{Y}}(y_j, y_i) = 0$, therefore $\lim_{r\to 0} \tilde{C}(i, j) = 0$.

(c) If $x_j - x_i > 0$ and $y_j - y_i < 0$, then $\lim_{r\to 0} L_{\mathbb{X}}(x_j, x_i) = 0$ and $\lim_{r\to 0} L_{\mathbb{Y}}(y_j, y_i) = 1$, therefore $\lim_{r\to 0} \tilde{C}(i, j) = 0$.

(d) If $x_j - x_i > 0$ and $y_j - y_i > 0$, then $\lim_{r\to 0} L_{\mathbb{X}}(x_j, x_i) = 0$ and $\lim_{r\to 0} L_{\mathbb{Y}}(y_j, y_i) = 0$, therefore $\lim_{r\to 0} \tilde{C}(i,j) = 1$.

Changing the index j by i in the previous reasoning for computing $\lim_{r\to 0} \tilde{C}(j,i)$, we conclude that if $x_j - x_i < 0$ and $y_j - y_i < 0$, then $\lim_{r\to 0} \tilde{C}(j,i) = 1$ and 0 in the rest of the cases. Analogously, for $\tilde{D}$, we have that $\lim_{r\to 0} \tilde{D}(i,j) = 1$ if $x_j - x_i > 0$ and $y_j - y_i < 0$ and 0 otherwise, and $\lim_{r\to 0} \tilde{D}(j,i) = 1$ if $x_j - x_i < 0$ and $y_j - y_i > 0$, and 0 in the rest of the cases. We end the proof by noting that

$$\lim_{r\to 0} \sum_{i=1}^{n} \sum_{j\neq i}^{n} \tilde{C}(i,j) = \sum_{i=1}^{n} \sum_{j\neq i}^{n} \lim_{r\to 0} \tilde{C}(i,j) = C$$
$$\lim_{r\to 0} \sum_{i=1}^{n} \sum_{j\neq i}^{n} \tilde{D}(i,j) = \sum_{i=1}^{n} \sum_{j\neq i}^{n} \lim_{r\to 0} \tilde{D}(i,j) = D.$$

ii. The proof is analogous to the previous one.

□

In principle, one may of course also look for the limits of $\tilde{\gamma}$ when $r \to \infty$. First, however, note that this case is hardly relevant from a practical point of view, as it means that all values are considered as completely tied. Theoretically, this case causes problems, too, since the limit does often not even exist. For example, when using E_r as an equivalence relation, it is easy to verify that $\tilde{T}(i,j) = 1$ for all i, j and hence $C = D = 0$ as soon as $r > 2 \max_i \{|x_i - x_i|, |y_i - y_j|\}$, which means that the numerator and the denominator in (12) is 0 and the term no longer well-defined.

There are, however, special cases in which the limit does indeed exist. Bodenhofer and Klawonn [8] point out that, in principle, the t-norm in (9–10) does not necessarily need to coincide with the t-norm underlying the

definition of the fuzzy order relations. For example, taking $\top = \min$ (and E_r as before), it can be seen that, for sufficiently large r, the degree of concordance (discordance) of each concordant (discordant) pair (x_i, y_i) and (x_j, y_j) is given by $r^{-1} \cdot \min\{|x_i - x_j|, |y_i - y_j|\}$. Thus, the parameter r in (12) cancels out, and the fuzzy gamma converges to

$$\tilde{\gamma} = \frac{\sum_{i<j}(c(i,j) - d(i,j)) \cdot m_{ij}}{\sum_{i<j}(c(i,j) + d(i,j)) \cdot m_{ij}} \ , \tag{13}$$

where $m_{ij} = \min\{|x_i - x_j|, |y_i - y_j|\}$, $c(i,j) = 1$ if (x_i, y_i) and (x_j, y_j) are concordant (and $c(i,j) = 0$ if not), and $d(i,j) = 1$ if (x_i, y_i) and (x_j, y_j) are discordant (and $d(i,j) = 0$ if not). In other words, $\tilde{\gamma}$ can be seen as a modification of the standard γ, in which the influence of each pair (x_i, y_i) and (x_j, y_j) is weighted by m_{ij}.

Despite the existence of the limit, (13) should be considered with caution, since the measure arguably loses its original character: Instead of considering closely neighbored data points as being tied to some degree, the idea of a tie loses its local property. Instead, the degree of concordance (discordance), and hence the degree of equivalence, are simply proportional to the dissimilarity of data points (as measured by the m_{ij}). Consequently, (13) is more numerical than rank-based (the concrete values x_i and y_i may have a strong influence) and partly loses its robustness properties. For example, consider a data set with observations $(x_1, y_1), \ldots, (x_{n+1}, y_{n+1})$, where $(x_i, y_i) = (i, i)$ for $i \leq n$ and $(x_{n+1}, y_{n+1}) = (M, -M)$ for some $M > n + 1$. Thus, while the first n values are perfectly (linearly) correlated, the last point is an outlier. The standard gamma is given by $\gamma = (n^2 - 3n)/(n^2 + n)$. Thus, it is robust and close to $+1$ for large n, regardless of M. In contrast, (13) strongly depends on the value of M, and even converges to -1 for $M \to \infty$.

4. Fuzzy Gamma as a Robust Correlation Measure

As a main motivation for their fuzzy extension of the gamma measure, the authors in [8] mention the goal to make the computation of rank correlation more robust toward noise in the data. In this section, we shall analyze the fuzzy gamma from this point of view. Thus, we assume that the observed data $\{(x_i, y_i)\}_{i=1}^n$ is corrupted with noise, which means that $x_i = \bar{x}_i + \epsilon_i$, where $\bar{x}_i$ is a true but unknown value, and ϵ_i is an error term independent of $\bar{x}_i$; likewise, $y_i = \bar{y}_i + \epsilon'_i$. As usual, the error terms are assumed to be independent and identically distributed.

We shall start with a kind of qualitative analysis of the effects of fuzzifying γ. Even though this analysis is based on some simplifying assumptions, it will help to develop a basic understanding of these effects. Moreover, it will be corroborated later one by means of suitable experiments.

It is fair to assume that adding random noise to the true data $\{(\bar{x}_i, \bar{y}_i)\}_{i=1}^n$ will resolve some of the existing ties, if any, but not create additional ones (if the error terms ϵ_i and ϵ'_i are real-valued random variables with a continuous density, then the probability to create a tie is indeed 0). So, with C_{true}, D_{true}, and T_{true} denoting, respectively, the *true* number of concordant, discordant, and tied pairs (among the $(\bar{x}_i, \bar{y}_i)$), and C_{obs}, D_{obs}, and T_{obs}, respectively, the *observed* number of concordant, discordant, and tied pairs (among the (x_i, y_i)), we have $T_{obs} \leq T_{true}$ or, equivalently, $C_{obs} + D_{obs} \geq C_{true} + D_{true}$. Moreover, since the distribution of error terms is typically symmetric with mean 0, it is natural to assume that a tie will be turned into a concordant

and discordant pair with equal probability, which means that

$$C_{obs} \approx C_{true} + \frac{1}{2}\left(T_{true} - T_{obs}\right) = C_{true} + \frac{\Delta T}{2} \ ,$$
$$D_{obs} \approx D_{true} + \frac{1}{2}\left(T_{true} - T_{obs}\right) = D_{true} + \frac{\Delta T}{2} \ ,$$

where $\Delta T = T_{true} - T_{obs} \geq 0$. Regarding the computation of rank correlation, we thus obtain

$$\gamma = \frac{C_{obs} - D_{obs}}{C_{obs} + D_{obs}} = \frac{C_{true} - D_{true}}{C_{true} + D_{true} + \Delta T} < \frac{C_{true} - D_{true}}{C_{true} + D_{true}} = \gamma_{true}$$

if $C_{true} > D_{true}$ and, likewise $\gamma > \gamma_{true}$ if $C_{true} < D_{true}$. In words, a computation of γ based on the observed data is biased toward 0, i.e., it will be an underestimation of truly positive and an overestimation of truly negative rank correlation coefficients.

Now, as mentioned previously, the basic principle underlying the fuzzy gamma measure is to turn concordant or discordant observations into partial ties. So, it can indeed be hoped that the "lost" ties ΔT will be recovered to some extent and, hence, that γ will be corrected in the right direction, namely toward the extreme values -1 or $+1$. Intuitively, it makes sense to assume that $\tilde{C} = (1 - \alpha)C_{obs}$, where α is the fraction of total concordance which is turned into equality via fuzzy equivalence. Obviously, this fraction depends on the scaling parameter r, so $\alpha = \alpha(r)$. Likewise, it makes sense to assume that $\tilde{D} = (1 - \alpha)D_{obs}$. In this case, however,

$$\tilde{\gamma} = \frac{(1-\alpha)C_{obs} - (1-\alpha)D_{obs}}{(1-\alpha)C_{obs} + (1-\alpha)D_{obs}} = \frac{C_{obs} - D_{obs}}{C_{obs} + D_{obs}} = \gamma \ , \tag{14}$$

i.e., $\tilde{\gamma}$ will be equal (or at least very similar) to γ. In other words, $\tilde{\gamma}$ is ineffective and does not correct γ toward γ_{true}.

It is important to note, however, that the above result is correct only if the fraction α is the same for the concordant and discordant pairs. Even though the concrete value of this quantity strongly depends on the data, it is interesting to note that the cardinality of the concordant and discordant pairs, respectively, is likely to have an important influence on this value. To explain this observation, assume that $C_{obs} > D_{obs}$, which means that the data is positively correlated; the case of negative correlation is treated analogously. While the distribution of uncorrelated data is typically a cloud having the same spatial extension in all direction, the distribution of positively correlated data is normally elongated, having the shape of a kind of ellipse; see Fig. 2 for an illustration. Now, suppose that a concordant and a discordant pair of data are picked at random. Under the above assumption of an elongated data distribution, the probability of being close to each other is higher for the discordant than for the concordant pair (there are many concordant pairs that are far from each other, but much less discordant pairs). This is confirmed by the cumulative distribution functions shown in Fig. 2.

These arguments imply that, in the case of positively correlated data, $\alpha < \beta$, where $\alpha = \alpha(r, C_{obs})$ is the fraction for concordant and $\beta = \beta(r, D_{obs})$ the fraction for discordant pairs that are turned into a tie. Consequently, (14) becomes

$$\tilde{\gamma} = \frac{(1-\alpha)C_{obs} - (1-\beta)D_{obs}}{(1-\alpha)C_{obs} + (1-\beta)D_{obs}} > \frac{C_{obs} - D_{obs}}{C_{obs} + D_{obs}} = \gamma \ ,$$

which means that $\tilde{\gamma}$ is indeed a proper correction of γ.

The above considerations provide evidence for the following conjectures:

- First, if the data to be analyzed is a noisy version of true data in which ties do exist, then the fuzzy gamma may potentially be a better

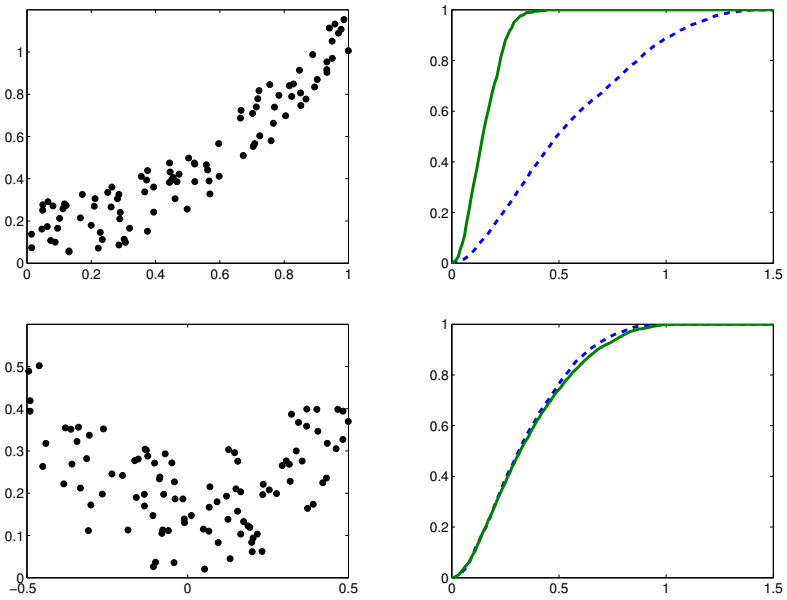

Figure 2: Typical examples of positively (top left) and uncorrelated (bottom left) data. The pictures to the right show a plot of the corresponding cumulative distribution functions mapping a distance d (x-axis) to the relative frequency of (concordant or discordant) data pairs whose distance is at most d (y-axis). The solid line depicts this function for the discordant data, the dashed one for the concordant data.

estimate than the original gamma; in this case, the performance of $\tilde{\gamma}$ will depend on the proper choice of the scaling parameter r.

- Second, if the original data does not contain any ties, then $\tilde{\gamma}$ is likely to give a biased estimate (while the original gamma is unbiased), as it will still tend to make the estimation more extreme (i.e., closer to $+1$ or -1).

- Third, if the original data does not contain any ties, but the observations are noisy, then $\tilde{\gamma}$ may potentially be a better estimate than the original gamma; like in the first case, the performance will depend on the proper choice of the scaling parameter r.

To explain the third conjecture, note that adding random noise to a data set will probably make the data less correlated, and the higher the level of noise, the stronger this effect will be (indeed, for a very high level of noise, the original data will be completely destroyed). Therefore, computing γ on the observed data will give a value which is probably closer to 0 than the true rank correlation, and since $\tilde{\gamma}$ tends to make the estimate more extreme, it might be able to compensate for this effect.

Since the second conjecture is actually a special case of the third one (noise level of 0), we may hypothesize that $\tilde{\gamma}$ can be beneficial whenever the original data is corrupted by noise, provided that the parameter r is chosen in a proper way (specifically, the second case calls for $r = 0$). In the following, we shall present some experimental studies to validate our conjectures.

4.1. Experiments

First evidence supporting our conjectures already comes from the experiments that have been conducted in [8]. In these experiments, synthetic data is produced by adding noise to a sample of points from the graph of a one-dimensional function. The first function $f_1(\cdot)$ used by the authors is piecewise linear with a big region of ties in the middle; see Fig. 3. The second function is parameterized and defined by $y = f_2(x) = x/2 + 1/4$ for $x \geq 0.5$ and $= (1 - 2q)x + q$ for $x \leq 0.5$. This function is monotone for $0 \leq q \leq 0.5$ and non-monotone for $0.5 \leq q \leq 1$; again, see Fig. 3.

Comparing the performance of $\tilde{\gamma}$ and γ as estimates of the rank correlation (which is $+1$ in the first two cases and close to 0 in the third), the authors find that $\tilde{\gamma}$ performs extremely well in the first case, comparatively good in the second case, but worse than γ in the third (non-monotone) case. These results are in complete agreement with our discussion above.

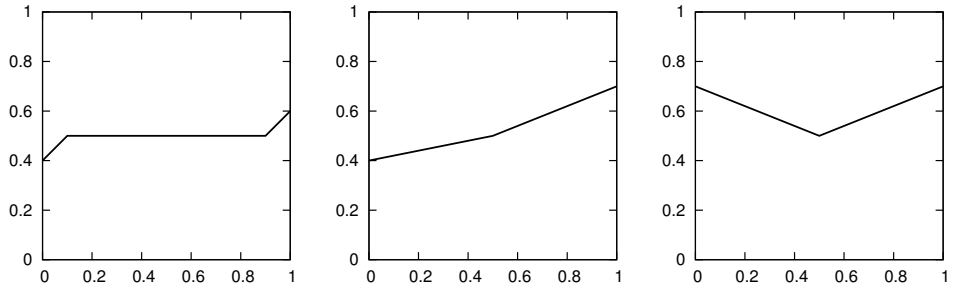

Figure 3: Graphs of the functions $f_1(\cdot)$ and $f_2(\cdot)$ for $q = 0.4$ (middle) and $q = 0.7$ (right).

To complement these experiments on synthetic data, we resorted to the well-known IRIS data set, a frequently used benchmark in data analysis.[3]

[3]For problems such as clustering and classification, this data set is actually not very

From the IRIS data, which comprises four real-valued variables and 150 observations. From the six possible two-dimensional combinations of the four features of the data set, we choose two representative ones. The first data set, D1, consists of the second and the fourth attribute, which are almost uncorrelated ($\gamma \approx -0.16$), and the second data set, D2, consists of the third and the fourth attribute, which are highly positively correlated ($\gamma \approx 0.84$); see Fig. 4. In D1 and D2, 14% and 9% of the data pairs are tied, respectively. We corrupted the data sets with random noise sampled from a normal distribution with mean 0 and standard deviation $\sigma \in \{0.008, 0.02, 0.04, 0.06, 0.1, 0.12, 0.15, 0.175, 0.2\}$. Moreover, we tried $\tilde{\gamma}$ with different values of the scaling parameter $r \in \{0.01, 0.06, 0.09, \ldots, 0.96\}$.

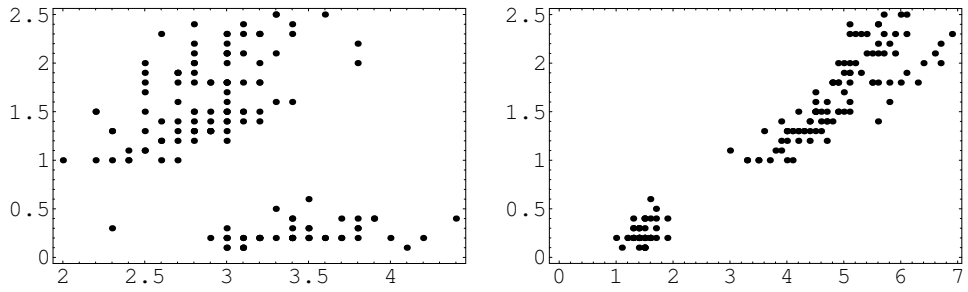

Figure 4: Left: Second and fourth attribute of the IRIS data ($\gamma \approx -0.16$). Right: Third and fourth attribute ($\gamma \approx 0.84$).

The results for data set D1 are shown in Fig. 5 ($\tilde{\gamma}$ as a function of the level of noise) and Fig. 6 ($\tilde{\gamma}$ as a function of r). In agreement with our expectations, $\tilde{\gamma}$ is not able to improve the estimation of γ. On the contrary, it tends to underestimate the true correlation, and the larger r, the stronger

challenging, which, however, is irrelevant for our purpose.

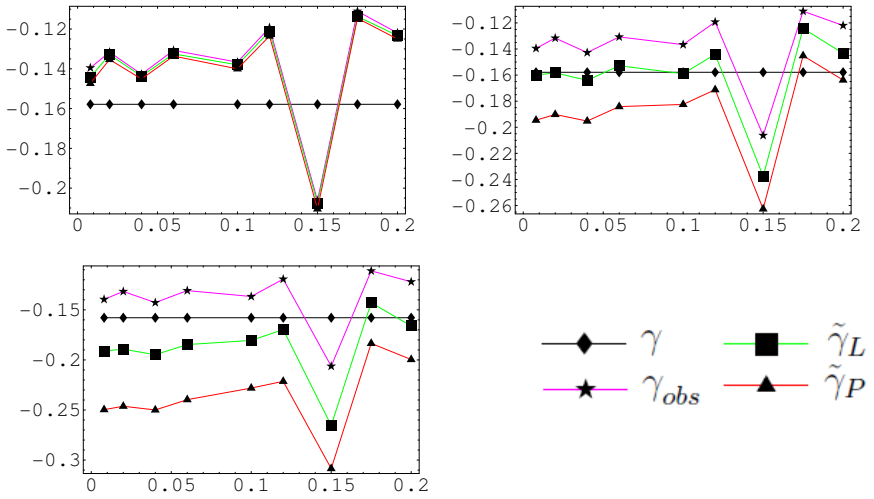

Figure 5: Rank correlation (y-axis) as a function of the level of noise (x-axis) for data set D1 and different values of r (left 0.01, middle 0.11, right 0.21)

.

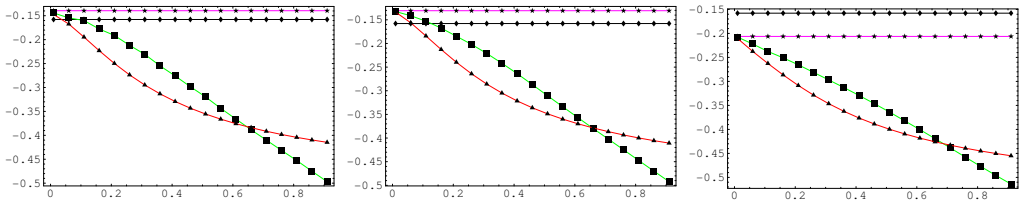

Figure 6: Rank correlation (y-axis) as a function of the scaling parameter r (x-axis) for data set D1 and different levels of noise (left 0.008, middle 0.06, right 0.175).

this effect becomes.

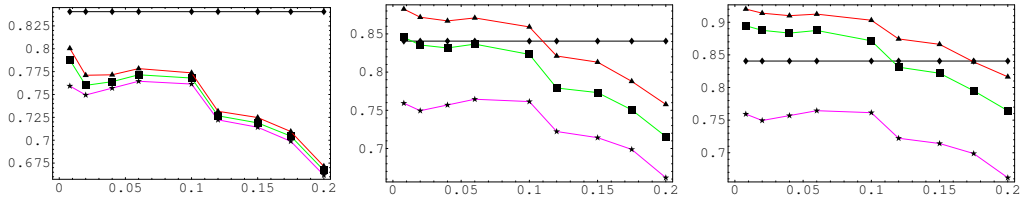

Figure 7: Rank correlation (y-axis) as a function of the level of noise (x-axis) for data set D2 and different values of r (left 0.01, middle 0.11, right 0.21).

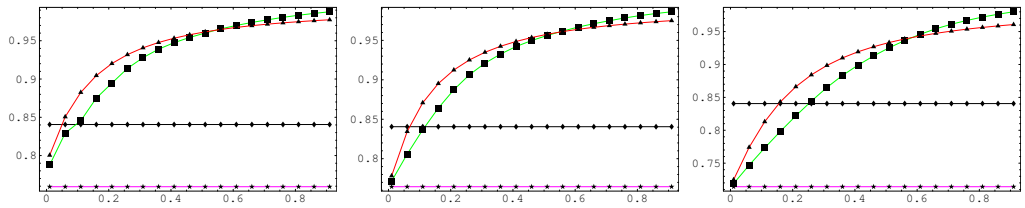

Figure 8: Rank correlation (y-axis) as a function of the scaling parameter r (x-axis) for data set D2 and different levels of noise (left 0.008, middle 0.06, right 0.15).

The results for data set D2 are shown in Fig. 7 and Fig. 8. This time, the performance of $\tilde{\gamma}$ is much better and again in agreement with what we expect. Indeed, $\tilde{\gamma}$ is able to improve the estimation of γ. In Fig. 8, it can nicely be seen that there is an optimal value of r which depends on the level of noise: The higher the noise, the larger r should be.

To validate our second and third conjecture, we need a data set without ties. To this end, we added a very small level of random noise to the data sets D1 and D2, respectively, and thus obtained two new data sets N1 and N2 without ties; see Fig. 9. For these data sets, which were now taken as the ground truth, we repeated the same experiments. The results, shown in Fig. 10 and Fig. 11 for N1 and in Fig. 12 and Fig. 13 for N2, are again in

agreement with our conjectures. If the noise added to N1 and N2, respectively, is very small, then $\tilde{\gamma}$ tends to give estimates biased toward -1 and $+1$ respectively. However, when the noise becomes larger, the original γ tends to give estimates biased toward 0, and $\tilde{\gamma}$ compensates for this, at least for a proper choice of r. An obvious example of this can be seen in the third case in Fig. 13, in which γ underestimates the true correlation of N2, and $\tilde{\gamma}_L$ ($\tilde{\gamma}_P$) repairs this for $r \approx 0.1$ ($r \approx 0.05$). Again, it can also be seen that higher levels of noise require higher values of r, i.e., the choice of the optimal r clearly depends on the level of noise.

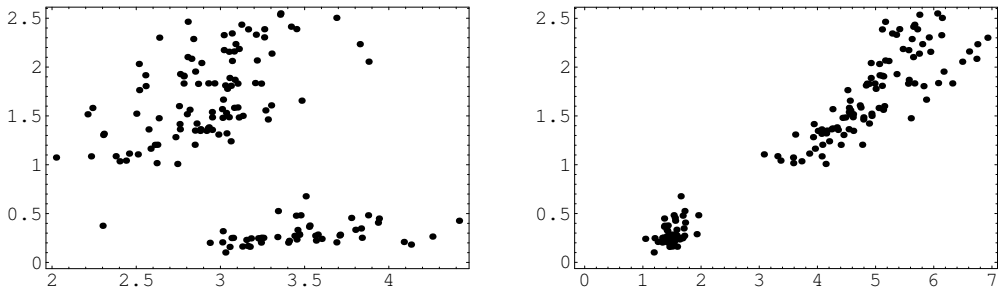

Figure 9: Left: Second and fourth attribute of noise-free data set N1 ($\gamma \approx -0.12$). Right: Third and fourth attribute of data set N2 ($\gamma \approx 0.77$).

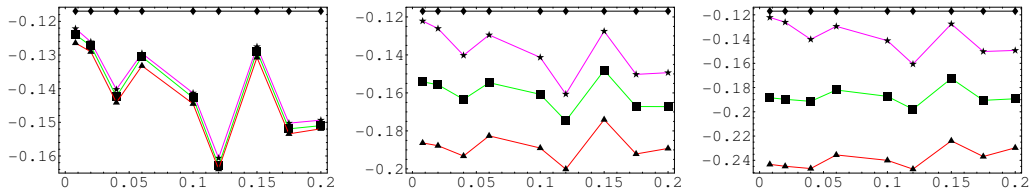

Figure 10: Rank correlation (y-axis) as a function of the level of noise (x-axis) for data set N1 and different values of r (left 0.01, middle 0.11, right 0.21).

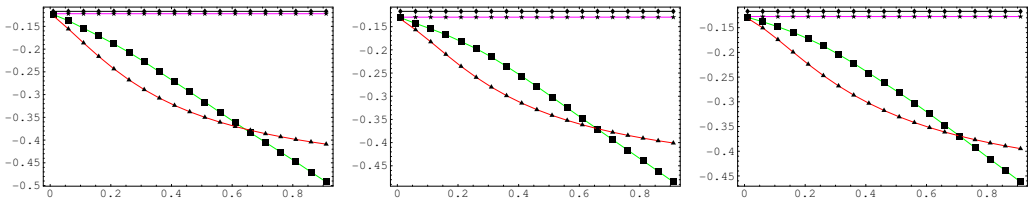

Figure 11: Rank correlation (y-axis) as a function of the scaling parameter r (x-axis) for data set N1 and different levels of noise (left 0.008, middle 0.06, right 0.15).

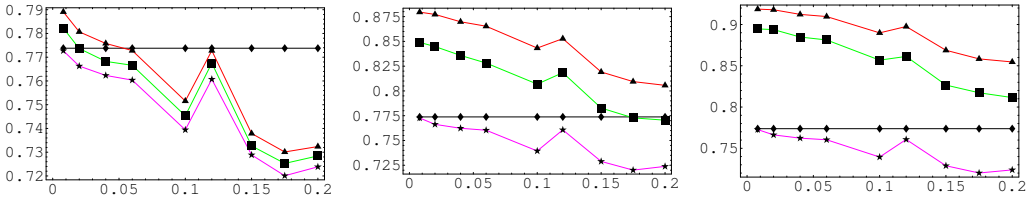

Figure 12: Rank correlation (y-axis) as a function of the level of noise (x-axis) for data set N2 and different values of r (left 0.01, middle 0.11, right 0.21).

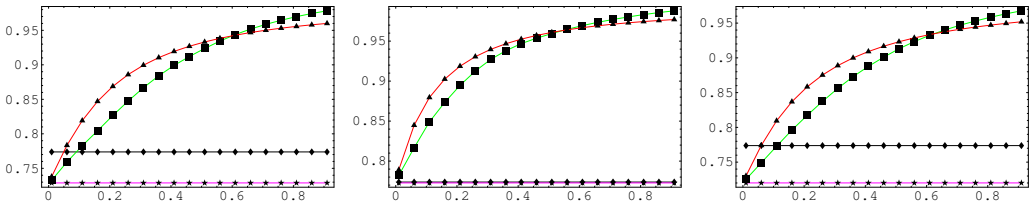

Figure 13: Rank correlation (y-axis) as a function of the scaling parameter r (x-axis) for data set N2 and different levels of noise (left 0.008, middle 0.06, right 0.15).

5. Perception-Based Rank Correlation

As mentioned earlier, the original motivation of the fuzzy gamma is to improve the estimation of rank correlation when data is corrupted with noise. This situation has been considered in the previous section. In this section, we propose an alternative and arguably not less interesting motivation. The idea is that, even though the data could in principle be observed without any errors, an overly precise measurement is actually not desired. In other words, even if two values x and y are precisely known and $x < y$, one may not want to distinguish between them, but instead treat them as being equal, at least to some extent.

In fact, in many situations, like in the example given in the introduction (correlation between article length and reviewer recommendation), very small differences are simply of no relevance. To given another example, the height of a person is normally measured up to a precision of one centimeter, which is completely sufficient, even though it could in principle be determined more precisely. However, putting two persons whose height differs by 1 mm on different ranks might simply not be desirable; instead, one may prefer to consider them as (almost) tied, which better agrees with human perception. Based on this idea, namely that the perceived differences are smaller than the actually measurable ones, we shall employ the term "perception-based rank correlation" for a measure that complies with the desired level of distinction.

The most straightforward way to realize a measure of this kind is to separate values into equivalence classes, i.e., to define an *equivalence relation* on the domain of an attribute. For numerical attributes, equivalence classes are reasonably chosen as intervals, which leads to an interval partition on a

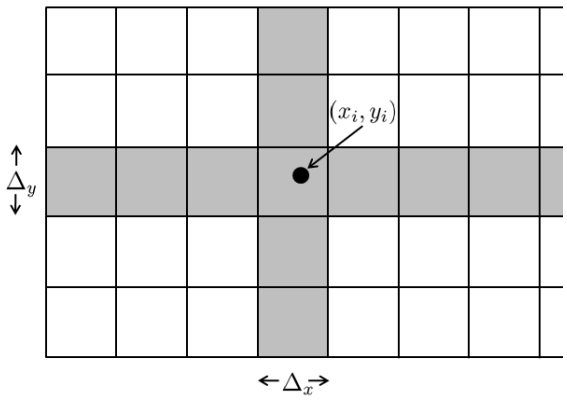

Figure 14: Illustration of the equivalence relation induced by a grid: All the points lying in the gray area are tied with the point (x_i, y_i).

one-dimensional domain and a two-dimensional grid in the case of a pair of variables; see Fig. 14.

Consider a one-dimensional partition defined by interval boundaries $\{\varepsilon_x \pm k \cdot \Delta_x \,|\, k \in \mathbb{Z}\}$, and another partition with boundaries $\{\varepsilon_y \pm k \cdot \Delta_y \,|\, k \in \mathbb{Z}\}$. Now, the idea is to consider two values within the same interval as being equal:

$$x_i \sim x_j \quad \Leftrightarrow \quad \exists\, k \in \mathbb{Z} : \varepsilon_x + k \cdot \Delta_x < x_i, x_j \leq \varepsilon_x + (k+1) \cdot \Delta_x \tag{15}$$

$$y_i \sim y_j \quad \Leftrightarrow \quad \exists\, k \in \mathbb{Z} : \varepsilon_y + k \cdot \Delta_y < y_i, y_j \leq \varepsilon_y + (k+1) \cdot \Delta_y \ . \tag{16}$$

With this definition of equality, a tuple (x_i, y_i), (x_j, y_j) is tied if it is located in the same "row" or the same "column" of the two-dimensional grid. Ignoring these ties, the gamma coefficient can be derived from the remaining tuples as usual. Subsequently, we shall refer to this coefficient as $\gamma(\Delta_x, \Delta_y, \varepsilon_x, \varepsilon_y)$.

A potential disadvantage of $\gamma(\varepsilon_x, \varepsilon_y, \Delta_x, \Delta_y)$ as defined above is its sensitivity toward the choice of the origin $(\varepsilon_x, \varepsilon_y)$. In fact, while Δ_x and Δ_y are in direct correspondence with the sought level of precision, the origin is often determined in a more or less arbitrary way. Obviously, the origin has an influence on the rank correlation through the determination of equivalence relations on $\mathbb{X}$ and $\mathbb{Y}$, respectively, a problem that is also known, for example, from the construction of histograms [14]. One idea to avoid this problem is to "average out" the origin, i.e., to derive the average of the gamma rank correlation over all origins. We call the resulting coefficient γ_{grid}:

$$\gamma_{grid} = \int_0^{\Delta_y} \int_0^{\Delta_x} \gamma(\Delta_x, \Delta_y, \varepsilon_x, \varepsilon_y)\, d\varepsilon_x d\varepsilon_y \tag{17}$$

For simplicity, and without loss of generality, we shall subsequently assume

$\Delta_x = \Delta_y = \Delta$. Note that, quite obviously, $\gamma_{grid} \to \gamma$ for $\Delta \to 0$; in fact, we have $\gamma_{grid} = \gamma$ as soon as $\Delta < \min_{1 \le i < j \le n} \min\{|x_i - x_j|, |y_i - y_j|\}$.

Just like the fuzzy gamma, γ_{grid} resorts to the idea of an equivalence relation on the underlying domains. In the case of γ_{grid}, however, this relation is non-fuzzy (i.e., it is a special case of a fuzzy equivalence relation underlying the fuzzy gamma). Intuitively, there should be a relationship between $\tilde{\gamma}$ and γ_{grid}, and one may expect that $\tilde{\gamma}$ in a sense mimics the averaging (17) over all non-fuzzy equivalence relations (indeed, note that $\tilde{\gamma}$ does not require the definition on any origin). In particular, one may expect that Δ is somehow in correspondence with the scaling parameter r in $\tilde{\gamma}$. In the remainder of this section, we shall investigate the relationship between $\tilde{\gamma}$ and γ_{grid} in more detail.

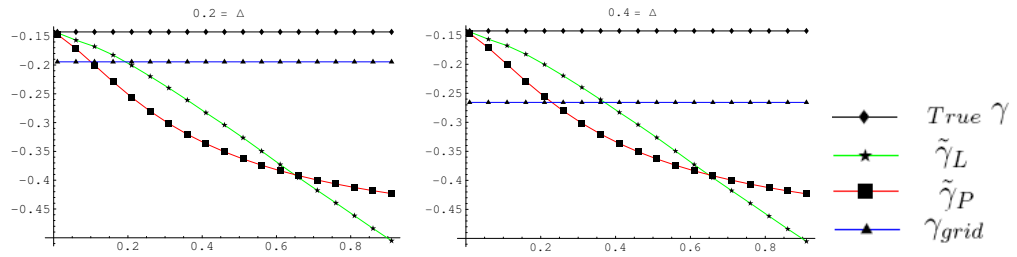

Figure 15: Rank correlation for data set N1, $\Delta = 0.2$ (left), $\Delta = 0.4$ (right). The x-axis corresponds to the values of r.

To get a first idea, we carried out some experiments with the data sets N1 and N2 from the previous section; see Fig. 9. For data set N1, Fig. 15 plots the values of the fuzzy gamma coefficients as a function of r, and compares them with γ_{grid} for $\Delta = 0.2$ and $\Delta = 0.4$, respectively. The same is shown in Fig. 16 for data set N2. As can be seen, we indeed have $\tilde{\gamma} \approx \gamma_{grid}$ if $r \approx \Delta$.

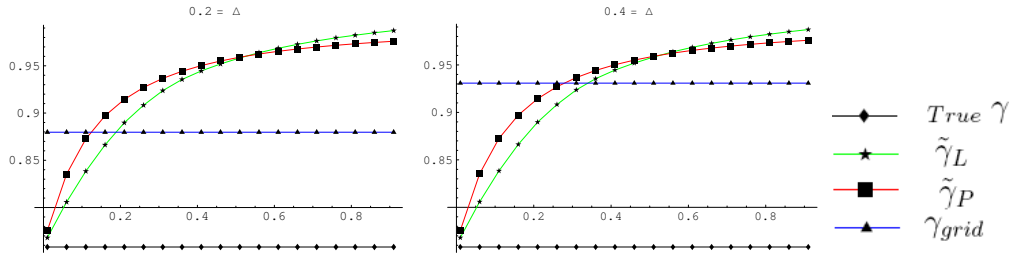

Figure 16: Rank correlation for data set N2, $\Delta = 0.2$ (left), $\Delta = 0.4$ (right). The x-axis corresponds to the values of r.

5.1. Relationship Between γ_{grid} and $\tilde{\gamma}$

In the following, we elaborate on the relationship between $\tilde{\gamma}$ and γ_{grid} in a more formal way. Replacing γ in (17) by its definition in terms of concordant and discordant pairs, we get

$$\gamma_{grid} = \int_0^{\Delta_y} \int_0^{\Delta_x} \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} (C(i,j) - D(i,j))}{\sum_{i=1}^{n} \sum_{j=i+1}^{n} (C(i,j) + D(i,j))} \, d\varepsilon_x d\varepsilon_y \qquad (18)$$

where the 0/1 variables

$$C(i,j) = C(\Delta_x, \Delta_y, \varepsilon_x, \varepsilon_y, i, j), \qquad D(i,j) = D(\Delta_x, \Delta_y, \varepsilon_x, \varepsilon_y, i, j)$$

indicate whether the index pair (i,j) is concordant or discordant, given the underlying grid specified by Δ_x, Δ_y, ε_x, ε_y:

$$C(i,j) = \begin{cases} 1 & \text{sign}(x_i - x_j) = \text{sign}(y_i - y_j) \text{ and } (x_i \not\sim x_j) \text{ and } (y_i \not\sim y_j) \\ 0 & \text{otherwise} \end{cases},$$

where $x_i \sim x_j$ and $y_i \sim y_j$ are defined according to (15) and (16), respectively; $D(i,j)$ is defined analogously. The definition of $T(i,j)$ then follows from $C(i,j) + D(i,j) + T(i,j) = 1$ (and is given by $T(i,j) = 1$ if $x_i \sim x_j$ or $y_i \sim y_j$ and $T(i,j) = 0$ otherwise).

Analyzing (18) is complicated by the denominator of the integrand, which corresponds to the number of ties obtained for the grid $(\Delta_x, \Delta_y, \varepsilon_x, \varepsilon_y)$. In the following, we make the simplifying and at least approximately valid assumption that this number is a constant.

Given our previous assumption on the number of ties, we get

$$\frac{\gamma_{grid}}{K} = \int_0^{\Delta_y} \int_0^{\Delta_x} \sum_{i=1}^{n} \sum_{j=i+1}^{n} C(i,j) - D(i,j)\, d\varepsilon_x d\varepsilon_y \tag{19}$$

where

$$K = \left(\sum_{i=1}^{n} \sum_{j=i+1}^{n} (C(i,j) + D(i,j)) \right)^{-1}$$

is constant because $\sum_{i=1}^{n} \sum_{j=i+1}^{n} (C(i,j)+D(i,j)) = \frac{1}{2}n(n-1) - \sum_{i=1}^{n} \sum_{j=i+1}^{n} T(i,j)$ and we have supposed that $\sum_{i=1}^{n} \sum_{j=i+1}^{n} T(i,j)$ remains constant regardless of the origin of the grid as specified by ε_x and ε_y. Using the linearity of the integral operator, the integrals in (19) can be moved inside the sums, and the expression can be rewritten as

$$\frac{\gamma_{grid}}{K} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \underbrace{\int_0^{\Delta_y} \int_0^{\Delta_x} C(i,j)\, d\varepsilon_x d\varepsilon_y}_{C_{grid}} - \sum_{i=1}^{n} \sum_{j=i+1}^{n} \underbrace{\int_0^{\Delta_y} \int_0^{\Delta_x} D(i,j)\, d\varepsilon_x d\varepsilon_y}_{D_{grid}} .$$

Thus, just like the other measures, γ_{grid} can be expressed as a function of the sum of pairwise degrees of concordance and discordance, respectively, above denoted by C_{grid} and D_{grid}:

$$C_{grid}(i,j) = \int_0^{\Delta_y} \int_0^{\Delta_x} C(i,j)\, d\varepsilon_x d\varepsilon_y, \tag{20}$$

$$D_{grid}(i,j) = \int_0^{\Delta_y} \int_0^{\Delta_x} D(i,j)\, d\varepsilon_x d\varepsilon_y \tag{21}$$

To compute these values, let (x_i, y_i) and (x_j, y_j) be a pair of points such that, without loss of generality, $(x_i, y_i) = (0, 0)$. We distinguish four cases:

Case 1. If $0 \leq x_j - x_i \leq \Delta_x$ and $0 \leq y_j - y_i \leq \Delta_y$ then

$$C_{grid}(i,j) = \int_{y_i}^{y_j} \frac{1}{\Delta_y} \left(\int_{x_i}^{x_j} \frac{1}{\Delta_x} d\varepsilon_x \right) d\varepsilon_y = \frac{(y_j - y_i)}{\Delta_y} \frac{(x_j - x_i)}{\Delta_x}$$

If $-\Delta_x \leq x_j - x_i \leq 0$ and $-\Delta_y \leq y_j - y_i \leq 0$, we can reason in a similar way. Thus, the results under these two conditions can be summarized as follows:

$$\begin{aligned} T_{grid}(i,j) &= 1 - \frac{|x_j - x_i|}{\Delta_x} \frac{|y_j - y_i|}{\Delta_y}, \\ C_{grid}(i,j) &= \frac{|x_j - x_i|}{\Delta_x} \frac{|y_j - y_i|}{\Delta_y}, \\ D_{grid}(i,j) &= 0. \end{aligned}$$

Case 2. If $0 \leq x_j - x_i \leq \Delta_x$ and $-\Delta_y \leq y_j - y_i \leq 0$ or if $-\Delta_x \leq x_j - x_i \leq 0$ and $0 \leq y_j - y_i \leq \Delta_y$, then

$$\begin{aligned} T_{grid}(i,j) &= 1 - \frac{|x_j - x_i|}{\Delta_x} \frac{|y_j - y_i|}{\Delta_y}, \\ C_{grid}(i,j) &= 0, \\ D_{grid}(i,j) &= \frac{|x_j - x_i|}{\Delta_x} \frac{|y_j - y_i|}{\Delta_y}. \end{aligned}$$

Case 3. If $0 \leq x_j - x_i \leq \Delta_x$ and $y_j - y_i \geq \Delta_y$ or if $-\Delta_x \leq x_j - x_i \leq 0$ and $y_j - y_i \leq -\Delta_y$, then

$$\begin{aligned} T_{grid}(i,j) &= 1 - \frac{|x_j - x_i|}{\Delta_x}, \\ C_{grid}(i,j) &= \frac{|x_j - x_i|}{\Delta_x}, \\ D_{grid}(i,j) &= 0. \end{aligned}$$

Case 4. If $0 \leq x_j - x_i \leq \Delta_x$ and $y_j - y_i \leq -\Delta_y$ or if $-\Delta_x \leq x_j - x_i \leq 0$ and $y_j - y_i \geq \Delta_y$, then

$$T_{grid}(i,j) = 1 - \frac{|x_j - x_i|}{\Delta_x},$$
$$C_{grid}(i,j) = 0,$$
$$D_{grid}(i,j) = \frac{|x_j - x_i|}{\Delta_x}.$$

The rest of the cases are straightforward and follow the same line of reasoning. Mering all these cases, we can thus express C_{grid}, D_{grid} and T_{grid} as follows:

$$T_{grid}(i,j) = 1 - \min\left(1, \frac{|x_j - x_i|}{\Delta_x}\right) \min\left(1, \frac{|y_j - y_i|}{\Delta_y}\right)$$

$$C_{grid}(i,j) = \begin{cases} \min\left(1, \frac{|x_j - x_i|}{\Delta_x}\right) \min\left(1, \frac{|y_j - y_i|}{\Delta_y}\right) & \text{if } \operatorname{sign}(x_j - x_i) = \operatorname{sign}(y_j - y_i) \\ 0 & \text{if } \operatorname{sign}(x_j - x_i) \neq \operatorname{sign}(y_j - y_i) \end{cases}$$

$$D_{grid}(i,j) = \begin{cases} \min\left(1, \frac{|x_j - x_i|}{\Delta_x}\right) \min\left(1, \frac{|y_j - y_i|}{\Delta_y}\right) & \text{if } \operatorname{sign}(x_j - x_i) \neq \operatorname{sign}(y_j - y_i) \\ 0 & \text{if } \operatorname{sign}(x_j - x_i) = \operatorname{sign}(y_j - y_i) \end{cases} \tag{22}$$

As can be seen from the above expressions, a comparison between values is done using an equivalence relation based on the Łukasiewicz t-norm, while these comparisons are then combined in terms of a product. Roughly speaking, γ_{grid} looks like a "hybrid" between $\tilde{\gamma}_L$ and $\tilde{\gamma}_P$. This impression is formally confirmed by the following proposition.

Proposition 2. γ_{grid} coincides with the fuzzy rank correlation $\tilde{\gamma}$ obtained by defining E_r in terms of the $\top_L$-equivalence (6) and using the product t-norm and conorm as aggregation operators in (9–11).

Proof. In order to see that $T_{grid}(i,j)$ is equivalent to $\tilde{T}(i,j)$ when using the equivalence relation E_r based on $\top_L$-equivalence and the product t-norm, note that

$$1 - \min\left(1, \frac{|x-y|}{\Delta}\right) = \max\left(0, 1 - \frac{|x-y|}{\Delta}\right) = E_r(x,y) \ ,$$

which is the $\top_L$-equivalence in equation (6) with $\Delta = r$. Furthermore,

$$T_{grid}(i,j) = 1 - \min\left(1, \frac{|x_i - x_j|}{\Delta}\right) \min\left(1, \frac{|y_i - y_j|}{\Delta}\right) = \perp_p(E_r(x_i,x_j), E_r(y_i,y_j))$$

where $\perp_p(x,y) = 1 - \top_p(1-x, 1-y) = 1 - (1-x)(1-y)$ and $\Delta = r$.

For the case of concordant and discordant pairs, it is enough to note that

$$\begin{aligned}
R_{\mathbb{X}}(x_i,x_j) &= 1 - L_{\mathbb{X}}(x_j,x_i) \\
&= 1 - \begin{cases} 0 & \text{if } x_j \leq x_i \\ E_r(x_j,x_i) & \text{otherwise} \end{cases} \\
&= \begin{cases} 0 & \text{if } x_j - x_i \leq 0 \\ 1 - \max(0, 1 - \frac{|x_j - x_i|}{r}) & \text{otherwise} \end{cases} \\
&= \begin{cases} 0 & \text{if } \operatorname{sign}(x_j - x_i) \text{ is negative} \\ \min(1, \frac{|x_j - x_i|}{r}) & \text{otherwise} \end{cases}
\end{aligned}$$

An equality of the same kind can be derived for $R_{\mathbb{Y}}(y_i,y_j)$. Thus, it is easy to see that

$$\begin{aligned}
C_{grid}(i,j) &= \top_P(R_{\mathbb{X}}(x_i,x_j), R_{\mathbb{Y}}(y_i,y_j)) = \tilde{C}(i,j) \ , \\
D_{grid}(i,j) &= \top_P(R_{\mathbb{X}}(x_i,x_j), R_{\mathbb{Y}}(y_j,y_i)) = \tilde{D}(i,j) \ ,
\end{aligned}$$

which proves the proposition. □

The above result shows that averaging the (non-fuzzy) grid-based measure over all origins of the grid yields a measure which is closely related to the idea of the fuzzy gamma coefficient. Indeed, as shown by (22), the concepts of concordance and discordance are fuzzified in exactly the same way, only by choosing a different combination of logical operators. By using the product instead of the Łukasiewicz t-norm as an aggregation operator, γ_{grid} achieves a somewhat smoother transition between concordance (discordance) and ties. This can be seen, for example, by comparing the tie-relations shown in Fig. 17. Still, one has to keep in mind that the above results are of an approximate nature, since all the derivations are based on the simplifying assumption of a constant denominator in (18). Thus, strictly speaking, γ_{grid} is not a sound fuzzy rank correlation from a theoretical point of view.

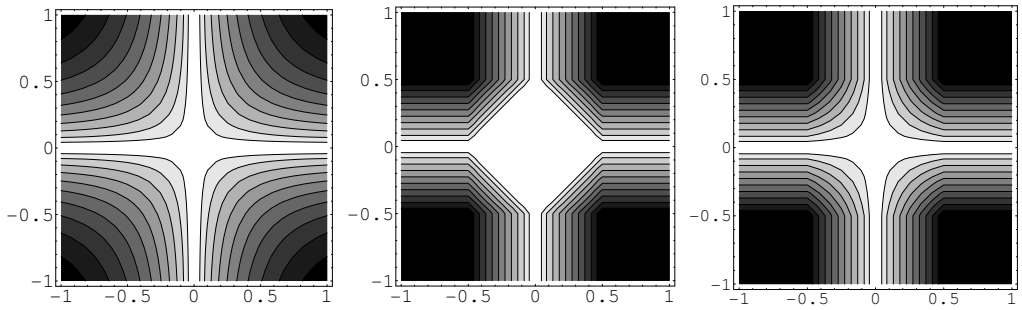

Figure 17: Contour plot of the tie-relation $\tilde{T}$ using the product (left) and Łukasiewicz (middle) t-norms. The right picture shows T_{grid}.

6. Concluding Remarks

In this paper, we have elaborated on a fuzzy extension of the well-known gamma rank correlation measure, which has recently been introduced by

Bodenhofer and Klawonn [8]. Apart from some minor technical points, the paper makes two major contributions:

- First, we corroborate the conjecture that the fuzzy gamma is advantageous in the presence of noisy data. More specifically, we offer formal arguments as well as empirical evidence for its ability to repair a bias of the original gamma, regardless of whether the true data contains ties or not.

- Second, we offer an alternative motivation of the fuzzy gamma in terms of a perception-based rank correlation measure and, in this regard, elaborate on its connection to a measure which proceeds from a non-fuzzy equivalence relation on the data space.

As to the first point, we already mentioned that a positive effect of the fuzzy gamma presumes a proper choice of the scaling parameter r. Although it was shown that this parameter is in direct correspondence with the level of noise in the data, the question of how to determine an optimal value for r was not addressed in this paper. Instead, this question is left for future work.

Another interesting question to be addressed in future work concerns the relation between fuzzy rank correlation and numeric correlation measures such as Pearson. In fact, one may also argue that a fuzzy rank correlation measure is somehow in-between a numeric and a purely rank-based measure. Therefore, it could possibly combine advantages from both sides. Finally, fuzzy rank correlation measures might be of interest in diverse fields of application, such as image processing, medicine, or bioinformatics, just to mention a few.

[1] J. Abrevaya. Computation of the maximum rank correlation estimator. *Economics Letters*, vol. 62, pp. 279-285, 1998.

[2] O. Ayinde and Y.H. Yang. Face recognition approach based on rank correlation of Gabor-filtered images, *Pattern Recognition*, vol. 35, pp. 1275-1289, 2002.

[3] R. Balasubramaniyan, E. Hüllermeier, N. Weskamp and J. Kämper. Clustering of Gene Expression Data Using a Local Shape-Based Similarity Measure, *Bioinformatics*, vol. 21, no. 7, pp. 1069-1077, 2005.

[4] D.N. Bhat and S.K. Nayar. Ordinal measures for image correspondence, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 4, pp. 415-423, 1998.

[5] U. Bodenhofer. A similarity-based generalization of fuzzy orderings preserving the classical axioms, *Int. Journal of Uncertainty, Fuzziness Knowledge-based Systems*, vol. 8, no. 5, pp. 593-610, 2000.

[6] U. Bodenhofer. Representations and constructions of similarity-based fuzzy orderings, *Fuzzy Sets and Systems*, vol. 137, pp. 113-136, 2003.

[7] U. Bodenhofer and M. Demirci. Strict Fuzzy Orderings with a Given Context of Similarity. Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 16, no. 2, pp. 147–178, 2008.

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

[9] T. Calders, B. Goethals and S. Jaroszewicz. Mining rank-correlated sets of numerical attributes, *In Proceedings of KDD'06*, Pennsylvania (USA), pp. 96-105, 2006.

[10] A. De Luca and S. Termini. A definition of non-probabilistic entropy in the setting of fuzzy sets theory. *Information and Control*, vol. 24, pp. 301-312, 1972.

[11] R. Fagin, R. Kumar and D. Sivakumar. Comparing top k lists, *In Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, Baltimore (Maryland), pp. 28-36, 2003.

[12] L.A. Goodman and W.H. Kruskal. *Measures of Association for Cross Classifications*, Springer-Verlag, New York, 1979.

[13] M. Kendall. *Rank Correlation Methods*, Charles Griffin & Company Limited, 1948.

[14] K. Loquin and O. Strauss. Histogram density estimators based upon a fuzzy partition. *Statistics & Probability Letters*, vol. 78, no. 13, 2008.

[15] M. Melucci. On rank correlation in information retrieval evaluation, *ACM SIGUR Forum*, vol. 41, n.1, pp. 18-33, 2007.

[16] V.J. Rayward-Smith. Statistics to measure correlation for data mining applications, *Computational Statistics & Data Analysis*, vol. 51, pp-3968-3982, 2007.

[17] Y. Shin. Rank estimation of monotone hazard models. *Economics Letters*, vol. 100, pp. 80-82, 2008.

[18] E. Yilmaz, J.A. Aslam and S. Robertson. A new rank correlation coefficient for information retrieval. *In Proceedings of SIGIR'08,* Singapore, pp. 587-594, 2008.