This is a pre- or post-print of an article published in
Bodenhofer, U., Krone, M., Klawonn, F.
Testing noisy numerical data for monotonic association
(2013) Information Sciences, 245, pp. 21-37.

Testing Noisy Numerical Data for Monotonic Association

Ulrich Bodenhofer[a,*], Martin Krone[b], Frank Klawonn[b,c]

[a]*Institute of Bioinformatics, Johannes Kepler University, 4040 Linz, Austria*
[b]*Department of Computer Science, Ostfalia University of Applied Sciences, 38302 Wolfenbüttel, Germany*
[c]*Bioinformatics and Statistics, Helmholtz Centre for Infection Research, 38124 Braunschweig, Germany*

Abstract

Rank correlation measures are intended to measure to which extent there is a monotonic association between two observables. While they are mainly designed for ordinal data, they are not ideally suited for noisy numerical data. In order to better account for noisy data, a family of rank correlation measures has previously been introduced that replaces classical ordering relations by fuzzy relations with smooth transitions — thereby ensuring that the correlation measure is continuous with respect to the data. The given paper briefly repeats the basic concepts behind this family of rank correlation measures and investigates it from the viewpoint of robust statistics. Then, on this basis, we introduce a framework of novel rank correlation tests. An extensive experimental evaluation using a large number of simulated data sets is presented which demonstrates that the new tests indeed outperform the classical variants in terms of type II error rates without sacrificing good performance in terms of type I error rates. This is mainly due to the fact that the new tests are more robust to noise for small samples. The Gaussian rank correlation estimator turned out to be the best choice in situations where no prior knowledge is available about the data, whereas the new family of robust gamma test provides an advantage in situations where information about the noise distribution is available. An implementation of all robust rank correlation tests used in this paper is available as an R package from the CRAN repository.

Keywords: gamma correlation coefficient, rank correlation, rank correlation test, fuzzy ordering, robust statistics, R package rococo

*Corresponding author

1. Introduction

Correlation measures are among the most basic tools in statistical data analysis and machine learning. They are applied to pairs of observations ($n \geq 2$)

$$(x_i, y_i)_{i=1}^{n} \tag{1}$$

to measure to which extent the two observations comply with a certain model. The most prominent representative is surely *Pearson's product moment coefficient* [1, 32], often called *correlation coefficient* for short. Pearson's product moment coefficient assumes a linear relationship as the underlying model.

Rank correlation measures [19, 26, 29] are intended to measure to which extent a monotonic function is able to model the inherent relationship between the two observables. They neither assume a specific parametric model nor specific distributions of the observables. Therefore, rank correlation measures are well-suited for detecting dependencies if no specific information about the data is available. The two most common approaches are *Spearman's rank correlation coefficient* (*Spearman's rho* for short) [34, 35] and *Kendall's tau (rank correlation coefficient)* [2, 25, 26]. Spearman's rho is defined as the Pearson product moment coefficient of the vectors of sorting ranks $(\text{rank}(x_1), \ldots, \text{rank}(x_n))$ and $(\text{rank}(y_1), \ldots, \text{rank}(y_n))$. The basic variant of Kendall's tau is defined as

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)},$$

where C and D are the numbers of concordant and discordant pairs, respectively:

$$C = |\{(i, j) \mid x_i < x_j \text{ and } y_i < y_j\}| \qquad D = |\{(i, j) \mid x_i < x_j \text{ and } y_i > y_j\}|$$

The rationale behind Kendall's tau is that every concordant pair counts as evidence for the assumption that the two observables are positively associated, whereas every discordant pair counts as evidence for a negative association between the observables. The more C exceeds D, the more likely a positive association is. Conversely, the more D exceeds C, the more likely a negative association is. Obviously, the presence of *ties*, i.e. pairs (i, j) for which either $x_i = x_j$ and/or $y_i = y_j$ holds, dilutes evidence in favor of an association between the two observables. For situations in which this is not desired, an advanced variant of Kendall's tau, commonly known as τ_b [26], and Goodman's and Kruskal's *gamma rank correlation measure* [19],

$$\gamma = \frac{C - D}{C + D},$$

have been introduced. Kendall's tau and γ coincide if there are no ties in the data. All rank correlation measures discussed are scaled to the interval $[-1, +1]$, are equal to $+1$ if there is a perfect positive association between the two observables and equal -1 in presence of a perfect negative association, whereas a value close to 0 indicates the absence of any monotonic association.

The rank correlation measures introduced above are perfectly suited for ordinal data, such as, ranks, marks, scores, etc., but there is vast number of applications in which numerical data must be tested for monotonic associations, e.g. in medical dose-response studies [18, 22, 24], in epidemiology [13], or when studying associations between gene expression levels [36]. In [5, 6], we argued in detail that the rank correlation measures mentioned above are not fully satisfactory for measuring rank correlation in numerical data that are perturbed by noise. Consequently, we proposed a family of rank correlation measures on the basis of fuzzy orderings. The present paper starts from this basis and introduces how a statistical test for monotonic association can be devised on the basis of this family of robust rank correlation measures. Subsequently, we present a detailed empirical evaluation of these rank correlation measures, also comparing them to the traditional rank correlation measures introduced above. This comparison also includes a recently published rank correlation measure for numerical data, the *Gaussian rank correlation estimator* [8]. It is conceptually similar to Spearman's rho, except for a monotonic transformation of the ranks. More specifically, the Gaussian rank correlation estimator is defined as the Pearson product moment correlation of the two vectors

$$\left(\Phi^{-1}\left(\tfrac{\mathrm{rank}(x_1)}{n+1}\right), \ldots, \Phi^{-1}\left(\tfrac{\mathrm{rank}(x_n)}{n+1}\right)\right) \quad \text{and} \quad \left(\Phi^{-1}\left(\tfrac{\mathrm{rank}(y_1)}{n+1}\right), \ldots, \Phi^{-1}\left(\tfrac{\mathrm{rank}(y_n)}{n+1}\right)\right),$$

where Φ is the cumulative distribution function of the standard normal distribution $\mathcal{N}(0, 1)$.

Testing for monotonic association is also closely related to testing for monotonicity of regression [9, 18, 20]. Our setup, however, is slightly different. First of all, we do not presume a model

$$Y = f(X) + \varepsilon$$

in which there is a designated "input observable" X and a designated "output observable" Y. Instead, we consider the two observables X and Y in a symmetric fashion like all standard correlation measures. Secondly, the null hypothesis of the tests for monotonicity of regression is that the underlying function f is monotonic. As it will be explained below, our null hypothesis is the independence of the two observables X and Y.

All considerations in this paper deal with standard (crisp) data and classical (crisp) p-values. We do not work with fuzzy data (e.g. [16, 31]), fuzzy estimates (e.g. [10]), or fuzzy p-values (e.g. [16, 17, 37]). From this point of view, our method applies fuzzy/soft techniques in classical statistics.

This paper is organized as follows. Section 2 highlights the family of robust rank correlation measures, whereas Section 3 introduces the corresponding statistical tests. Then Section 4 briefly introduces an open-source software package that implements the robust rank correlation measures and the corresponding tests introduced in this paper. Section 5 then relates the robust rank correlation measures to concepts from robust statistics in order to formally argue in favor of the robustness of our family of rank correlation measures. An empirical study that compares robust rank correlation tests with their classical counterparts is presented in Sections 6 and 7. Section 6 compares type II error rates for given significance thresholds, whereas Section 7 compares type I error rates.

2. The Family of Robust Rank Correlation Measures

The rank correlation measures introduced above have the advantage that they are able to detect monotonic association without making any assumption concerning the distribution of the data and the specific kind of association. Therefore, they are ideal for detecting monotonic associations in ordinal data. If numerical data are to be considered, however, the issue of noise sensitivity arises. Random perturbations of the data, even if they are small, may obscure monotonic association. We have argued in detail in [5, 6] that all classical rank correlation measures are non-continuous with respect to the data and that small random perturbations, i.e. noise, may severely impair the ability to detect monotonic association. Consequently, we have introduced a family of rank correlation measures that allow for continuity with respect to the data. The basic idea behind this family is to replace the strict orderings in the definitions of concordant and discordant pairs by continuous functions that measure the *degree to which one value is greater than another*. In the following, we will consider *non-negative scoring functions* $R : X^2 \to \mathbb{R}$, where X is a linearly ordered set, having the following properties:

(i) Irreflexivity: $R(x, x) = 0$
(ii) Monotonicity: $x \leq y$ implies $R(x, z) \leq R(y, z)$ and $R(z, x) \geq R(z, y)$

Note that these two properties further entail $R(y, x) = 0$ whenever $x \leq y$ and therefore, asymmetry, i.e. $\min(R(x, y), R(y, x)) = 0$.

In principle, it is possible to use any binary (two-argument) scoring function that is non-decreasing in one and non-increasing in the other component. In our preceding papers [5, 6], we decided in favor of using *strict fuzzy orderings* [3, 4] as scoring functions, since a solid mathematical grounding is available for them. Since this is not of primary relevance for this paper, the reader is referred to [5, 6].

Let us briefly mention some examples of scoring functions on the real numbers that fulfill the above-mentioned properties:

ε-tolerant strict ordering: (for $\varepsilon \geq 0$)

$$R_\varepsilon^{\text{crisp}}(x, y) = \begin{cases} 1 & \text{if } y > x + \varepsilon \\ 0 & \text{otherwise} \end{cases}$$

This definition also includes the classical strict ordering for $\varepsilon = 0$.

Truncated linear scoring: (with parameter $r > 0$)

$$R_r^{\text{lin}}(x, y) = \min\big(\max\big(\tfrac{1}{r}(y - x), 0\big), 1\big)$$

This scoring function can be interpreted as follows: the degree to which x is smaller than y is 0 if $x > y$ and 1 if $y \geq x + r$ (compare with $R_\varepsilon^{\text{crisp}}$ above). For $y \in [x, x + r]$, the degree to which y is greater than x increases linearly from 0 to 1. The parameter r can be interpreted as a *radius of tolerance*.

Laplace scoring: (with parameter $b > 0$)

$$R_b^{\text{exp}}(x, y) = \max\big(1 - \exp\big(-\tfrac{1}{b}(y - x)\big), 0\big)$$

Gaussian scoring: (with parameter $\sigma > 0$)

$$R_\sigma^{\text{Gauss}}(x, y) = \begin{cases} 1 - \exp(-\frac{1}{2\sigma^2}(x - y)^2) & \text{if } y > x \\ 0 & \text{otherwise} \end{cases}$$

All these four examples have in common that $R(x, y) \in [0, 1]$. Except the first one, all are continuous in both arguments. All except the fourth scoring function can be interpreted as strict fuzzy orderings in the sense of [4].

For a given scoring function $R : X^2 \to [0, 1]$, we can define an operator $E : X^2 \to [0, 1]$ as

$$E(x, y) = 1 - \max(R(x, y), R(y, x)).$$

For the four scoring functions above, this results in the following:

$$E_{\varepsilon}^{\text{crisp}}(x,y) = \begin{cases} 1 & \text{if } |x-y| \leq \varepsilon \\ 0 & \text{otherwise} \end{cases}$$
$$E_{r}^{\text{lin}}(x,y) = \max(1 - \tfrac{1}{r}|x-y|, 0)$$
$$E_{b}^{\text{exp}}(x,y) = \exp\big(-\tfrac{1}{b}|x-y|\big)$$
$$E_{\sigma}^{\text{Gauss}}(x,y) = \exp(-\tfrac{1}{2\sigma^2}(x-y)^2)$$

All these four operators can be interpreted as some underlying similarity scoring functions. Obviously, E_b^{exp} and E_σ^{Gauss} are unnormalized densities of Laplace and Gaussian distributions, respectively (which is the reason why we named R_b^{exp} and R_σ^{Gauss} in this way above).

Now suppose that we are given pairs of real-valued data as in (1) and that we have chosen appropriate scoring functions $R_X : \mathbb{R}^2 \to [0,1]$ and $R_Y : \mathbb{R}^2 \to [0,1]$ for the first and the second observable, respectively. Given an index pair (i, j), we can compute the degree to which (i, j) is a concordant pair as

$$\tilde{C}(i,j) = \bar{T}(R_X(x_i, x_j), R_Y(y_i, y_j))$$

and the degree to which (i, j) is a discordant pair as

$$\tilde{D}(i,j) = \bar{T}(R_X(x_i, x_j), R_Y(y_j, y_i)),$$

where $\bar{T}$ is some binary function used for aggregating the relationships of x and y components in a conjunctive manner. Then we can compute the overall score of concordant pairs $\tilde{C}$ and the overall score of discordant pairs $\tilde{D}$, respectively, as sums of the following scores:

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

Consequently, we can define the *generalized gamma rank correlation measure* $\tilde{\gamma}$ as

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

We speak of a *family of rank correlation measures* because the concrete realization of $\tilde{\gamma}$ depends on the choice of R_X, R_Y, and $\bar{T}$. In our previous works, we assumed $\bar{T}$ to be a triangular norm [27]. For this paper, we can restrict to three simple choices: $T_{\mathbf{M}}(x,y) = \min(x,y)$, $T_{\mathbf{P}}(x,y) = x \cdot y$ and $T_{\mathbf{L}}(x,y) = \max(x+y-1, 0)$.

An important property of $\tilde{\gamma}$ is immediate to see: if the scoring functions R_X and R_Y are continuous and if $\bar{T}$ is continuous too, then $\tilde{\gamma}$ depends continuously on the data set $(x_i, y_i)_{i=1}^n$. In [5, 6], we provided empirical arguments in favor of the robustness of the family of generalized gamma rank correlation coefficients. A more formal analysis of mathematical properties is provided in [33]. In this paper, we introduce tests based on this family of rank correlation coefficients (cf. Section 3) and we will also demonstrate the robustness of the measures in Section 5. Therefore, from now on, we will speak of "robust gamma rank correlation coefficients/tests".

3. Rank Correlation Tests Based on the Family of Robust Rank Correlation Coefficients

Association measures can be applied in various ways, e.g. for mining dependencies between observables, for unsupervised grouping of features or supervised feature selection. Our family of robust rank correlation measures has also been applied in such a way already [28]. The most prominent application, however, is to use them in a statistical test to determine whether there is a significant association in the data. For the classical rank correlation measures, such tests have been available for a long time and have become standard tools in statistics. The main contribution of this paper is to introduce association tests for our family of robust rank correlation measures. As usual in correlation tests, the null hypothesis $\mathcal{H}_0$ is that the two observables under consideration are independent, whereas the alternate hypothesis $\mathcal{H}_1$ is that the two observables are dependent. Analogously to well-established correlation tests, we distinguish between one-sided tests (for positive or negative association) and the two-sided test for any association. The procedure is the same for all these tests: we first compute the test statistic $\tilde{\gamma}$ on the given data and then, assuming the null hypothesis of independence, compute the (approximate) probability of having observed this value or a more extreme one. If this probability (the well-known *p-value*) is smaller than a given *significance level* α, the null hypothesis is rejected, otherwise the null hypothesis is accepted.

Computing the exact probability of having observed a certain value of $\tilde{\gamma}$ or a more extreme one under the null hypothesis requires full knowledge of the test statistic's distribution under the null hypothesis (often called the *null distribution*). No such explicit distributions are known for the classical gamma or Kendall's tau. Our rank correlation coefficient further involves the scoring functions R_X and R_Y as well as the aggregation function $\bar{T}$, hence, the problem of determining the exact distribution becomes even more difficult. That is why it is more than unlikely

that an explicit representation of the null distribution for our robust rank correlation coefficients can be determined. Instead, we resort to the well-known remedy of *permutation testing*: assuming that the two observables are independent, any combination of two values is equally likely. Therefore, we can obtain all possible values of the test statistic for the given data under the null hypothesis by considering all permutations of one of the two observables. More specifically, if we compute the test statistic $\tilde{\gamma}$ for $\mathbf{x} = (x_1, \ldots, x_n)$ and every possible permutation of $\mathbf{y} = (y_1, \ldots, y_n)$, then we can compute the p-values as follows:

Test for positive association: p = relative frequency of how many times the test statistics for the shuffles are at least as large as $\tilde{\gamma}$ for the original, unshuffled data;

Test for negative association: p = relative frequency of how many times the test statistics for the shuffles are at most as large as $\tilde{\gamma}$ for the unshuffled data;

Two-sided test: p = relative frequency of how many times the absolute value of the test statistics for the shuffles is at least as large as the absolute value of $\tilde{\gamma}$ for the unshuffled data;

This approach has the advantage that we need not have an analytic representation of the null distribution. For the classical rank correlation measures introduced in Section 1, considering all permutations would provide exact p-values. The same is true for the above procedures, but only conditional to the fact that the samples are fixed. Without this assumption, however, the null distribution also depends on the marginal distributions of the two observables which are not necessarily known. So, depending on how representative the data are for the underlying marginal distributions, the p-values may be more or less imprecise.

In cases in which it is intractable or too expensive to consider all permutations, we can only resort to computing an approximation of the p-value. In principle, there are two ways of doing this: As a first method, we can create a certain number m of *random shuffles* of $\mathbf{y}$ and count the number of times that the test statistic was at least as extreme as the test statistic on the unshuffled data. This number S is binomially distributed with m trials and probability p, where p is the exact p-value, and we can derive the approximate p-value as $\frac{S}{m}$. Again under the assumption that the data are fixed, $\frac{S}{m}$ is an unbiased estimate of the true p-value.

To approximate the null distribution by some known distribution would be an alternative to the method above. This variant is commonly used for Kendall's rank correlation test in which a normal distribution is used to approximate the null distribution. The parameters of this normal distribution fortunately depend only on

the number of samples n. In the case of our family of robust rank correlation coefficients, this method cannot be used. The null distribution for our robust gamma rank correlation coefficients, on the one hand, depends on the choice of $\bar{T}$ and the two scoring functions R_X and R_Y. On the other hand, it does not only depend on the ordering of the data vectors $\mathbf{x}$ and $\mathbf{y}$, but also on the marginal distributions of the two observables. Experiments suggest (see Supplementary Figure S1) that the null distributions are close to normal. However, in most experiments described in Section 6, all well-established tests for normality rejected the normality assumption in a majority of cases. Therefore, we do not further pursue this option.

Once the p-value of a test for given data is computed, we can apply a significance threshold α, i.e., we accept the null hypothesis if $p \geq \alpha$ and reject the null hypothesis if $p < \alpha$. If the p-values were exact, the *type I error rate*, i.e., the probability of wrongly rejecting the null hypothesis, also called *false positive rate (FPR)*, would be exactly α. However, as noted above, the p-values are not exact, and they are not even exact for classical correlation tests or only under additional assumptions. That is why we will investigate type I error rates empirically (see Section 7).

4. The R Package rococo

We have implemented all tests introduced in this paper as an R package named `rococo` [7]. This package is available publicly and freely through the *Comprehensive R Archive Network (CRAN)*[1]. Further instructions can be found at `http://www.bioinf.jku.at/software/rococo/`.

The package implements a wide selection of robust rank correlation coefficients. The user is free to choose R_X and R_Y from the four parametric families of scoring functions mentioned in Section 2, namely $R_\varepsilon^{\text{crisp}}$, R_r^{lin}, R_b^{exp}, and R_σ^{Gauss}. For the aggregation operator $\bar{T}$, three choices are available: the minimum t-norm $T_{\mathbf{M}}$, the product t-norm $T_{\mathbf{P}}$, and the Łukasiewicz t-norm $T_{\mathbf{L}}$. The user can even supply custom functions for aggregation, however, at the cost of a severe slowdown.

Most importantly, the package implements rank correlation tests based on these robust rank correlation coefficients. All tests return a p-value based on permutation testing as described above and, additionally, a p-value based on a normal approximation of the null distribution. For ten or less samples, it is possible to consider the complete set of $n!$ permutations. Additionally, the package also implements the Gaussian rank correlation estimator [8].

[1] `http://cran.r-project.org/web/packages/rococo/`

5. Robustness Properties

Robust statistics (see, for instance, [21, 23, 30]) emphasizes the importance of handling outliers and limiting the influence of single data and noise on the outcome of a statistical analysis. In this section, we briefly recall the basic concepts from robust statistics and compare our family of robust rank correlation coefficients to the established measures from Section 1.

5.1. Breakdown Point

The *breakdown point* of a statistic is usually defined as the largest possible fraction of samples for which the statistic is still bounded when that fraction is altered without restriction [21]. This definition is not useful in the context of correlation coefficients, since they only yield values between -1 and 1, so an unbounded change can never occur. Therefore, we follow [8] and define the breakdown point as the smallest fraction of arbitrary contamination needed to make the correlation coefficient uninformative, i.e. to change the sign of the correlation coefficient. Of course, the breakdown point depends on the sample. For reasons of comparability with other rank correlation measures that are based on ranks, we investigate the breakdown point with respect to the identical sample $\mathbf{s}_n = \{(1,1), \ldots, (n,n)\}$; hence, we have $\mathbf{x} = (1, \ldots, n)$ and $\mathbf{y} = (1, \ldots, n)$. Trivially, $\mathbf{s}_n$ yields a correlation of 1 according to all classical correlation measures, and to our new variants too, since the value of $\tilde{D}$ will always be zero. In this setting, the breakdown point of the Pearson correlation coefficient is always 0, whereas Spearman's rho and Kendall's tau have limit breakdown points (if n goes to infinity) of 0.206 and 0.293, respectively [8, 12]. For the Gaussian rank correlation estimator, the limit breakdown point has been determined as 0.124 [8].

The breakdown points of the robust rank correlation coefficients discussed in this paper depend on the choice of the underlying strict fuzzy ordering, its parameter and the chosen t-norm. In the following, we determine the breakdown point for all combinations of orderings and t-norms that are available in the `rococo` package (see Section 4). For the analysis, we will assume that the same strict ordering is used, i.e. $R_X = R_Y$.

5.1.1. Breakdown Point for the Classical Strict Ordering

In case of the classical strict ordering R_0^{crisp}, the values of R_X and R_Y are either 0 or 1. As $\bar{T}(0,0) = \bar{T}(0,1) = \bar{T}(1,0) = 0$ and $\bar{T}(1,1) = 1$ holds for every t-norm $\bar{T}$, the values of $\tilde{C}$ and $\tilde{D}$ correspond to the number of concordant and discordant pairs in Kendall's tau. Thus, the breakdown point equals the breakdown point of Kendall's tau (i.e. 0.293).

5.1.2. Breakdown Point for Fixed Parameter

Except for the classical strict ordering, every ordering available in the `rococo` package can be parameterized by a value larger than zero. We assume that this parameter is fixed and is used for both orderings. In this setting, the values of R_X and R_Y are either zero or, for $n \to \infty$, an infinite proportion of them will tend to 1 whereas only a fixed proportion will be between zero and one. Thus, $\tilde{C}$ and $\tilde{D}$ will converge to the number of concordant and discordant pairs in Kendall's tau and the limit breakdown point will again correspond to the breakdown point of Kendall's tau.

5.1.3. Breakdown Point for the Parameter Depending on the Interquartile Range

In the `rococo` implementation, the value of the parameter is automatically chosen as 20% of the interquartile ranges of $\mathbf{x}$ and $\mathbf{y}$, respectively, if this parameter is not supplied by the user. The interquartile range of a sample $\mathbf{t} = (t_1, \ldots, t_n)$ is defined as $q(\mathbf{t}, 0.75) - q(\mathbf{t}, 0.25)$ with q being the sample quantile function

$$q(\mathbf{t}, \alpha) = t_{(i)} \quad \text{if } (i-1)/n < \alpha \leq i/n, \quad i = 1, \ldots, n\,,$$

where $t_{(1)}, \ldots, t_{(n)}$ is derived by sorting $\mathbf{t}$ in ascending order.

In order to determine the breakdown point if the parameter is chosen as 20% of the interquartile ranges, the outliers that result in the largest decrease of our rank correlation coefficients are sought. Due to the similarity of the proposed measures and Kendall's tau, the greatest effect that k outliers have is obtained if $\mathbf{y} = (1, \ldots, n)$ is changed to

$$\mathbf{y}' = (y'_1, y'_2, \ldots, y'_n)$$

in such a way that $y'_1 > y'_2 > \ldots > y'_k > y'_n$ where $y'_i = i$ for $i = k+1, \ldots, n$ is unchanged. With the help of order statistics, this can be written as $y'_{(i)} = y'_{k+i}$ for $i = 1, \ldots, n-k$ and $y'_{(n+1-j)} = y'_j$ for $j = 1, \ldots, k$. In the following, we will commonly refer to the parameter of the ordering as r_x (for $\mathbf{x}$) and $r_{y'}$ (for $\mathbf{y}'$), although the parameter is denoted ε for the ε-intolerant strict ordering, r for the truncated linear scoring, b for the Laplace scoring and σ for the Gaussian scoring. To visualize the influence of the outliers on the value of $\tilde{\gamma}$, we use two matrices $M_{\mathbf{x}}$ and $M_{\mathbf{y}'}$ with $M_{\mathbf{x},i,j} = R_{X,r_x}(x_i, x_j)$ and $M_{\mathbf{y}',i,j} = R_{Y,r_{y'}}(y'_i, y'_j)$. In this setting, the value of $\tilde{C}$ can be calculated by applying the t-norm $\bar{T}$ to every corresponding pair of entries of $M_{\mathbf{x}}$ and $M_{\mathbf{y}'}$, whereas for $\tilde{D}$, the transpose of $M_{\mathbf{y}'}$ has to be used.

In order to determine the breakdown points of our rank correlation coefficients, the cases $k < 0.25n$ and $k \geq 0.25n$ need to be considered separately.

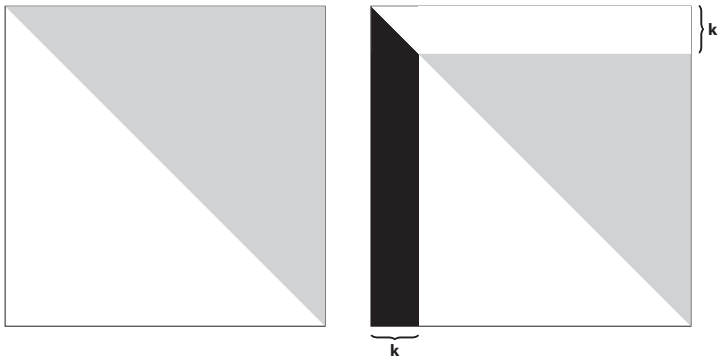

Figure 1: Form of the matrices $M_{\mathbf{x}}$ (left) and $M_{\mathbf{y}'}$ (right).

- For $k < 0.25n$, the interquartile range of $\mathbf{y}'$ is not influenced by the outliers and hence $r_x = r_{y'} = \frac{n}{10}$. In order to decrease the value of the rank correlation coefficients as much as possible, the pairwise differences $y'_i - y'_{i+1}$ for $i = 1, \ldots, k$ need to be chosen such that $M_{\mathbf{y}',i,j}$, $j = 1, \ldots, k$ and $i = j + 1, \ldots, n$, is as close to 1 as possible. In case of the truncated linear scoring, for example, it is sufficient to ensure that the pairwise differences are at least as large as $r_{y'}$. Figure 1 describes the matrices $M_{\mathbf{x}}$ and $M_{\mathbf{y}'}$ for $\mathbf{y}'$ having k outliers subject to the above considerations. In this figure, the black areas correspond to values arbitrarily close to one, whereas the entries inside the white areas are zero. Within the grey areas, the values range between 0 and 1 with entries further away from the main diagonal being closer to 1. In this setting, the overall score of concordant pairs is

$$\tilde{C}_{n,k} = \sum_{r=k+1}^{n} \sum_{c=r+1}^{n} \bar{T}\left(R_{X,r_x}(r,c), R_{Y,r_{y'}}(r,c)\right)$$

(recall that $x_i = i$ for $i = 1, \ldots, n$ and $y'_i = i$ for $i = k+1, \ldots, n$), whereas the overall score of discordant pairs equals

$$\tilde{D}_{n,k} = \sum_{r=1}^{k} \sum_{c=r+1}^{n} R_{X,r_x}(r,c) \ ,$$

since the corresponding entries in the transpose of $M_{\mathbf{y}'}$ will be 1. The one-sided limit

$$\lim_{a \to 0.25^-} \left(\lim_{\substack{n \to \infty \\ k = \lfloor a \cdot n \rfloor}} \frac{\tilde{C}_{n,k} - \tilde{D}_{n,k}}{\tilde{C}_{n,k} + \tilde{D}_{n,k}} \right) \tag{2}$$

Table 1: Values of (2) for the orderings and t-norms in the `rococo` package if the parameter of the ordering is chosen as 20% of the interquartile range.

	$T_{\mathbf{M}}$	$T_{\mathbf{P}}$	$T_{\mathbf{L}}$
R_r^{lin}	47/542	1/16	26/521
R_b^{exp}	0.05485	< 0	< 0
R_σ^{Gauss}	0.0264	< 0	< 0
$R_\varepsilon^{\text{crisp}}$	7/162	7/162	7/162

(i.e. for a approaching 0.25 "from below") was calculated for the combinations of ordering and t-norm available in the `rococo` package. The obtained values can be found in Table 1 and an explanation is given in Supplementary Section S1. Clearly, for those combinations of ordering and t-norm where the limit is less than zero, the breakdown point is less than 0.25. In these cases, we used numerical simulations to obtain the breakdown point. The breakdown points for R_b^{exp} are 0.244 and 0.234 in case of the product t-norm $T_{\mathbf{P}}$ and the Łukasiewicz t-norm $T_{\mathbf{L}}$, respectively. For R_σ^{Gauss}, the breakdown points are 0.239 for $T_{\mathbf{P}}$ and 0.231 for $T_{\mathbf{L}}$. In the other cases, the breakdown point of our proposed robust rank correlation coefficient is at least 0.25.

- We will now consider only those combinations of ordering and t-norm for which the limit in Table 1 is larger than zero. For $k \geq 0.25n$, the interquartile range of $\mathbf{y}'$ is no longer independent of the outliers. Consequently, we have $r_x = \frac{n}{10}$, whereas

$$r_{y'} = \begin{cases} \frac{1}{5}\left(y'_{\frac{n}{4}} - \frac{3(k+1)+n}{4}\right) & \text{if } k < 0.75n \\ \frac{1}{5}\left(y'_{\frac{n}{4}} - y'_{\frac{3n}{4}}\right) & \text{otherwise.} \end{cases}$$

  Consequently, $r_{y'}$ directly depends on the values of the outliers.

  The largest decrease of $\tilde{\gamma}$ in case of $k \geq 0.25n$ is obtained if the two following conditions are fulfilled.

  1. $\tilde{C}$ needs to be as small as possible, i.e. $R_{Y,r_{y'}}(y'_i, y'_j)$ for $i = k+1, \ldots, n$ and $j = i+1, \ldots, n$ has to be as close as possible to 0.
  2. The value of $\tilde{D}$ has to be as large as possible. Hence, $R_{Y,r_{y'}}(y'_i, y'_j)$ for $j = 1, \ldots, k$ and $i = j+1, \ldots, n$ needs to be as close as possible to 1.

  The first condition is fulfilled if $r_{y'}$ is as large as possible. For the second condition, it is neccessary for the pairwise differences $y'_i - y'_{i+1}$ for

$i = 1, \ldots, k$ to be very large. In case of the truncated linear scoring, for example, these differences need to be at least as large as $r_{y'}$. In this setting, the overall score of concordant pairs is arbitrarily close to zero, whereas the overall scrore of discordant pairs is as large as

$$\sum_{i=1}^{k}(n-i) = \frac{1}{2}\left(2kn - k^2 - k\right) .$$

Hence, if the outliers are chosen in accordance to the above considerations, the value of $\tilde{\gamma}$ is arbitrarily close to -1. Consequently, the breakdown point for those combinations of ordering and t-norm for which the value of (2) is larger than zero is exactely 0.25 if the parameter of the ordering is chosen as 20% of the interquartile ranges.

5.2. Sensitivity Curve

The *sensitivity curve* is another important concept of robust statistics. It shows the influence of a single new observation added to a sample. In the context of correlation coefficients, a new observation is a pair of values (x, y), hence, the sensitivity curve is actually a function of two variables. The sensitivity "curve" of a correlation coefficient ρ for a sample $\mathbf{z}_{n-1} = \{(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})\}$ is defined as (cf. [8])

$$\mathrm{SC}_n(x, y; \rho, \mathbf{z}_{n-1}) = n \cdot (\rho(\mathbf{z}_{n-1} \cup \{(x, y)\}) - \rho(\mathbf{z}_{n-1})) .$$

For a robust correlation coefficient, the sensitivity curve should be bounded, i.e. the influence of a single new observation should be limited. Figure 2 shows two sensitivity curves for Kendall's tau. For the left curve, 10 pairs of samples from two independent uniform distributions on the unit interval were used, having a correlation of almost 0. The right curve is based on the identical sample of size 10, i.e. $\mathbf{s}_{10} = \{(1, 1), \ldots, (10, 10)\}$. We observe that the sensitivity curves of Kendall's tau are not continuous, but piecewise constant.

Figure 3 shows two sensitivity curves of our robust rank correlation coefficients for the uncorrelated sample on which the sensitivity curve on the left-hand side of Figure 2 is based. Both sensitivity curves are based on $R_X = R_Y = R_r^{\mathrm{lin}}$ with $r = 0.1$ (left) and $r = 0.5$ (right). For aggregation, $\bar{T} = T_{\mathbf{L}}$ has been used. In contrast to the sensitivity curves of Kendall's tau, the sensitivity curves of the robust rank correlation coefficient are continuous and it can be clearly seen that, the larger r, the smoother the sensitivity curve. Continuous sensitivity curves are

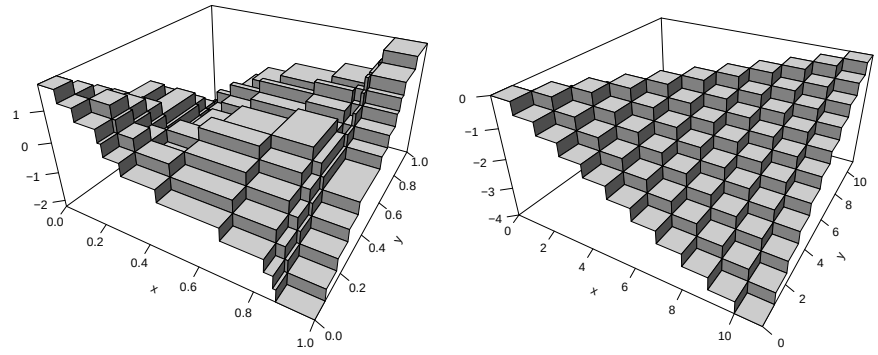

Figure 2: Two sensitivity curves for Kendall's tau for samples with $n = 10$. Left: sensitivity curve for random sample generated from two independent uniform distributions on the unit interval; right: sensitivity curve for perfectly correlated sample.

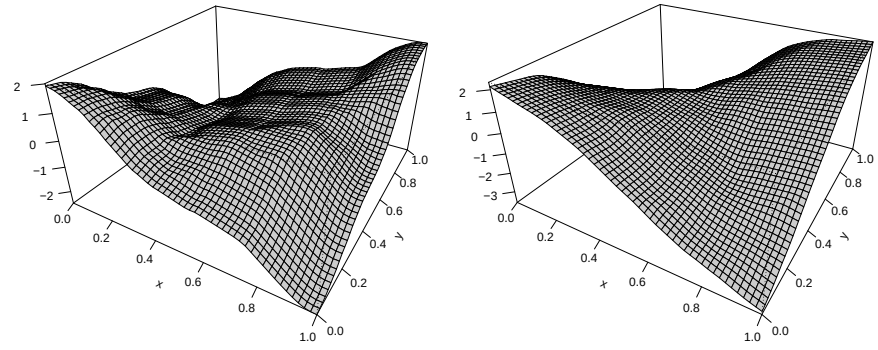

Figure 3: Two sensitivity curves for the robust rank correlation coefficient based on the fuzzy ordering R_r^{lin} for a sample created from independent uniform distributions. Left: $r = 0.1$; right: $r = 0.5$.

advantageous since a slight change in the values of the new pair (x, y) will only have a small impact on the value of the rank correlation coefficient.

The *finite sample gross error sensitivity* of a correlation coefficient ρ is the worst case scenario for the sensitivity curve for the perfectly correlated sample $\mathbf{s}_n$

$$\text{FSGE}_n(\rho) = \sup_{x,y\in\mathbb{R}} |\text{SC}_n(x, y; \rho, s_{n-1})| \tag{3}$$

(cf. [8]). Clearly, for a robust correlation coefficient, the influence of a single new observation should be limited, i.e. the finite sample gross error sensitivity should be as small as possible. Again, for reasons of comparability, the analysis is done

with respect to the identical sample of size n, i.e. $\{(1,1),\ldots,(n,n)\}$. Thus, the finite sample gross error sensitivity of the Pearson correlation coefficient is $2n$, whereas it is 4 for Kendall's tau and 6 for Spearman's rho [8]. For the Gaussian rank correlation estimator, the finite sample gross error sensitivity is proportional to $2\log(n)$ (cf. [8]). This implies that the *asymptotic gross error sensitivity*, i.e., $\lim_{n\to\infty} \mathrm{FSGE}_n(\rho)$, of both the Spearman correlation coefficient and Kendall's tau is constant, whereas it is unbounded for both the Pearson correlation coefficient and the Gaussian rank correlation estimator. The asymptotic gross error sensitivity of our robust rank correlation coefficients again depends on the choice of the underlying strict fuzzy ordering, its parameter and the t-norm. For the analysis, we again assume that the same ordering is used, i.e. $R_X = R_Y$.

5.2.1. Asymptotic Gross Error Sensitivity for the Classical Strict Ordering

In case of the classical strict ordering R_0^{crisp}, the maximum influence of a single new observation (x_n, y_n) is obtained for $x_n > \max_i(x_i)$, i.e. $x_n > n-1$, and $y_n = -x_n$. Clearly, the choice of the t-norm is irrelevant and $\tilde{C}$ equals $\frac{1}{2}(n^2-3n+2)$, whereas $\tilde{D} = n-1$. Hence, we have $\tilde{\gamma} = 1 - \frac{4}{n}$ and both the finite sample gross error sensitivity and the asymptotic gross error sensitivity are 4, i.e. they coincide with those of Kendall's tau.

5.2.2. Asymptotic Gross Error Sensitivity for Fixed Parameter

We will now consider the asymptotic gross error sensitivity in the case that the parameter of the chosen ordering is fixed. Then, the largest influence of a single new observation (x_n, y_n) is still obtained by taking $y_n = -x_n$. However, the value of x_n needs to be chosen depending on the underlying strict ordering and its parameter. In this setting, the value of $\tilde{D}$ is as large as $n-1$. Due to the complex structure of our robust rank correlation coefficients, we succeeded in obtaining closed formulae for the value of $\tilde{C}$ in the expression $\tilde{\gamma}\,(s_{n-1} \cup \{(x_n, y_n)\})$ for only a limited number of combinations of ordering and t-norm. These formulas are listed in Supplementary Section S2. In the other cases, we used numerical computations to obtain the asymptotic gross error sensitivity. We observed that if the parameter of the ordering is fixed, the asymptotic gross error sensitivity is always 4, i.e. it coincides with the asymptotic gross error sensitivity of Kendall's tau. Furthermore, the smaller the value of the parameter is, the faster the value of (3) tends to this limit.

5.2.3. Asymptotic Gross Error Sensitivity for the Parameter Depending on the Interquartile Range

If the parameter of the ordering is chosen as 20% of the interquartile ranges of **x** and **y**, its value is not influenced by the single new observation (x_n, y_n), provided that $n \geq 5$. Consequently, the value of the parameter is $\frac{n-1}{10}$. Again, we used numerical computations to determine the asymptotic gross error sensitivity in these settings and it turns out that the limits are larger than the limit for Kendall's tau, but smaller than the limit for Spearman's rho. The exact values we obtained can be found in Supplementary Table S1.

The above considerations show that the proposed robust rank correlation coefficients have good robustness properties. Although, depending on how the parameter of the ordering is chosen, the robustness properties may be worse than those of Kendall's tau, the proposed measures have been shown to be more robust to noise than Spearman's rho. Furthermore, the robust rank correlation coefficients have the additional advantage of having continuous sensitivity curves.

6. Experimental Validation of Type II Error Rates

Since the complex structure of the robust rank correlation coefficient does not facilitate a rigorous analytical investigation, we have to resort to an empirical study in which we compare the results of several rank correlation coefficients on a large number of simulated data sets.

6.1. Data Sets

We considered a total number of 2,880 random data sets, all of which were created according to the model

$$y = f(x) + \varepsilon, \tag{4}$$

i.e. the second observable y is given by applying a function f to the first observable x plus additive independent noise ε.

We considered three sizes of data sets, $n = 10$, $n = 20$, and $n = 40$. For creating values of the first observable, two different distributions were considered: (a) standard normal distribution with mean 0 and variance 1 and (b) a symmetric uniform distribution of values between -2 and 2. For the function underlying the model (4), we considered four different families of functions:

1. Identity: $f_1(x) = x$;

2. Cubic polynomial: $f_2(x) = \frac{x^3}{4}$; the scaling with $\frac{1}{4}$ has been chosen to obtain values approximately in the same range as the other functions.
3. Piecewise linear function with flat area:

$$f_3(x) = \begin{cases} x + a & \text{if } x \leq -a \\ x - a & \text{if } x \geq a \\ 0 & \text{otherwise} \end{cases}$$

   The parameter a is chosen randomly according to a uniform distribution of values between 0.2 and 1.
4. Piecewise linear function with small decreasing part:

$$f_4(x) = \begin{cases} x + 2a & \text{if } x \leq -a \\ x - 2a & \text{if } x \geq a \\ -x & \text{otherwise} \end{cases}$$

   The parameter a is chosen randomly according to a uniform distribution of values between 0.1 and 0.3.

We consider the following noise distributions: normally distributed noise with $\mu = 0$ and three different choices of $\sigma \in \{0.05, 0.1, 0.25\}$ as well as Laplacian noise with $\mu = 0$ and three different choices of $b \in \{0.1, 0.2, 0.5\}$. So, in total, six different noise distributions are used. For each combination of settings listed above, 20 replicates were created, which results in a total of 2,880 data sets:

$$\underbrace{2}_{\text{input dist.}} \cdot \underbrace{3}_{\text{sample sizes}} \cdot \underbrace{4}_{\text{functions } f} \cdot \underbrace{6}_{\text{noise dist.}} \cdot \underbrace{20}_{\text{replicates}} = 2880.$$

For each data set involving functions f_3 or f_4, the parameter a is chosen independently. Figure 4 shows 12 plots of examples of data sets along with their underlying function f.

6.2. Compared Rank Correlation Tests

We applied seven rank correlation tests to all 2,880 data sets. For comparison with our robust rank correlation tests, we used the Spearman test, the Kendall test, and the classical gamma test. We did not include the Pearson test, as this test is only suitable for linearly correlated data and is sensitive to outliers [8]. All three included tests solely depend on the ranking of the values in the data set and are, therefore, independent of the scaling of the data and their marginal distributions. From our new framework, we tested the following variants:

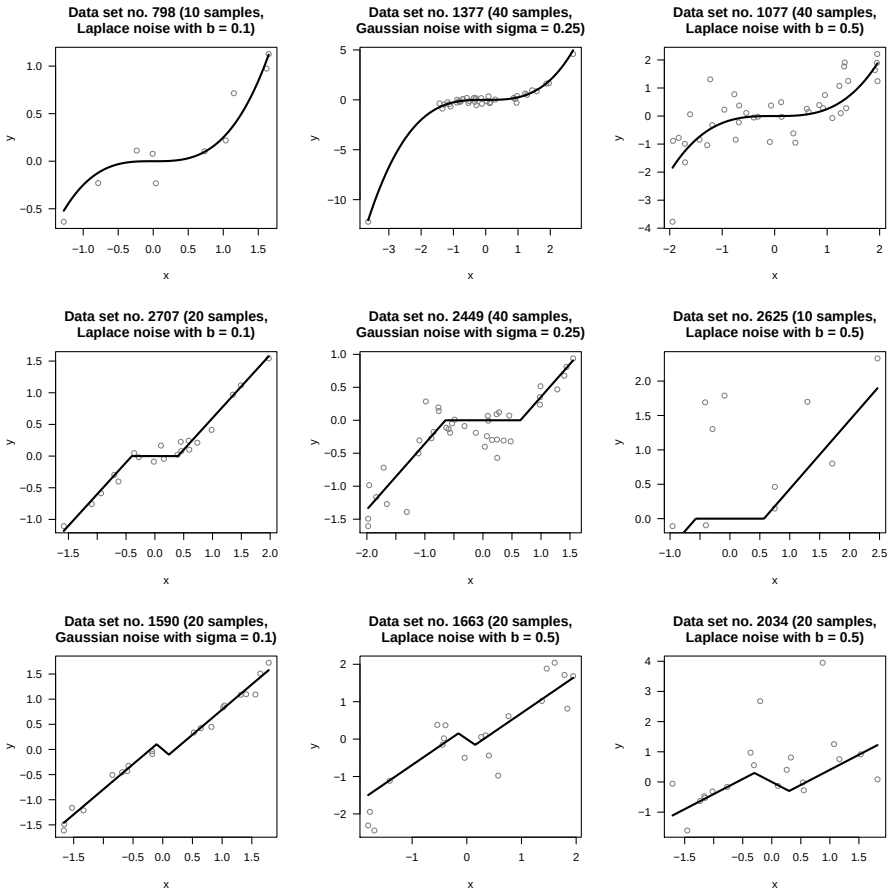

Figure 4: Nine of the 2,880 simulated data sets along with their exact model functions; the first row shows data sets that were created using function f_2, the second row shows data sets created with f_3, while the third row depicts data sets created using f_4.

No prior knowledge: suppose we have no knowledge about the model underlying the data. Then an ad-hoc choice of parameters has to be made. We used $R_X = R^{\text{lin}}_{r_x}$ and $R_Y = R^{\text{lin}}_{r_y}$, i.e. strict fuzzy orderings based on triangle similarities for both components, where the tolerance radii r_x and r_y were chosen as 20% of the interquartile ranges of the values $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$, respectively. We will refer to this variant as *robust gamma test with ad-hoc settings* in the following.

Using prior knowledge: suppose we want to test whether our data have been generated from the model (4) for an *unknown monotonic function* f and a

known noise distribution. If we want to make use of this prior information, it is immediate to use the following settings:

- $R_X = R_0^{\text{crisp}}$, i.e. we use the crisp strict ordering for the first component, as there is no noise on the first component in the model (4).
- If ε is normally distributed with mean 0 and standard deviation σ, use $R_Y = R_\sigma^{\text{Gauss}}$.
- If ε is Laplace-distributed with mean 0 and scale parameter b, use $R_Y = R_b^{\text{exp}}$.

We will refer to this variant as *robust gamma test using prior information*.

As a third variant, we consider the case that we again know that the data comply with the model (4) for an unknown monotonic function f and an unknown noise distribution, however, with known standard deviation σ. One possibility to tackle this situation is to use the following settings:

- $R_X = R_0^{\text{crisp}}$ (as above);
- $R_Y = R_\sigma^{\text{crisp}}$, i.e. crisp strict ordering not taking any pairs into account whose difference in the y component is less than the noise level.

We will refer to this third variant as *noise-tolerant gamma test*.

The robust gamma tests with ad-hoc settings were made using $\bar{T} = T_{\mathbf{M}}$. In the other two tests, at least one of the two scoring functions is binary, therefore, it does not matter whether we use $\bar{T} = T_{\mathbf{M}}$, $\bar{T} = T_{\mathbf{P}}$, or $\bar{T} = T_{\mathbf{L}}$.

Finally, we included the Gaussian rank correlation estimator [8] in the comparison.

Knowing that all functions underlying our data are largely non-decreasing, we performed tests for positive association only. All analyses were done using the statistical computing platform R. For the Spearman test and the Kendall test, the standard R function `cor.test()` was used. The classical gamma test, the three robust rank correlation tests, and the Gaussian rank correlation test were performed using our R package `rococo` (see Section 4). For data sets with 10 samples, all permutations were considered. For data sets with 20 or 40 samples, approximate tests were performed using 100,000 random permutations (see Section 3).

Supplementary Figure S1 shows histograms of rank correlation values (according to the ad-hoc settings) for random permutations of the twelve data sets

shown in Figure 4. It can be observed that, the larger the sample, the more the values of $\tilde{\gamma}$ seem to be normally distributed. For 10 samples, the histograms show the largest deviations from a normal distribution, whereas for 40 samples, there is virtually no deviation of the histograms from the fitted normal distribution. It seems that the null distributions converge to normal distributions. However, for most of our data sets, well-established normality tests rejected the normality hypothesis. The asymptotic null distribution would require further investigation. Since this is not the main focus of this paper, we defer this to future studies.

6.3. Evaluation Results

The main question about a statistical test is whether it correctly accepts or rejects the null hypothesis. The data sets in this section have all been created such that the two observables are positively correlated. So the type II error rates on these data corresponds to the proportion of falsely accepted null hypotheses. Table 2 shows the type II error rates for the seven tests using the four most popular significance thresholds 0.05, 0.01, 0.005, and 0.001. The table provides these rates for all 2,880 data sets together and also split up by the monotonic functions underlying the model (4). Regardless of the threshold and the type of underlying model, the three robust tests and the Gaussian rank correlation test clearly outperform the three classical tests. Of the four robust tests, the one that makes use of prior information works best, followed by the Gaussian rank correlation test.

The results above rely on the choice of a particular significance level α. The question arises whether we can assess the performance of the tests regardless of a particular choice of a certain significance level. The simplest approach is to plot the type II error rates (FNR) versus all possible choices of α. If we plot TPR $= 1 -$ FNR versus α instead, the resulting curve can be interpreted as a receiver-operator characteristic (ROC) curve [15]. Figure 5 shows these curves for all seven tests we compared. In order to emphasize the differences between the curves, we restrict to the most illustrative top-left corner of the curves (FPR $\in [0, 0.06]$, TNR $\in [0.875, 1]$). If we make a vertical cut at a certain significance level α, we can directly see the type II error rate as one minus the true negative rate (compare the cuts in Figure 5 with the values in Table 2). The ROC curves confirm the results of Table 2: all four robust tests outperform all three classical tests, where the test with optimal prior information again performs best, again followed by the Gaussian rank correlation test.

The standard measure for evaluating an ROC curve is the area under the curve (AUC). Table 2 also provides such an evaluation for the seven compared tests.

Table 2: Type II error rates (false negative rates) of the seven compared tests with respect to four different significance levels. The first block gives cumulated rates over all 2,880 data sets. The other four blocks correspond to rates computed for sub-selections of data sets created with a specific model function f.

			Spearman	Gaussian	Kendall	gamma	robust gamma (ad-hoc settings)	robust gamma (using prior inf.)	robust gamma (noise-tolerant)
all	FNR at $\alpha =$	0.05	0.0403	0.0326	0.0458	0.0458	0.0375	0.0312	0.0354
		0.01	0.0941	0.0750	0.0851	0.0854	0.0795	0.0677	0.0774
		0.005	0.1201	0.0941	0.1080	0.1080	0.1021	0.0882	0.1052
		0.001	0.1924	0.1476	0.1965	0.1969	0.1646	0.1500	0.1722
	1 – AUC		0.0093	0.0077	0.0099	0.0099	0.0085	0.0074	0.0087
f_1 (linear)	FNR at $\alpha =$	0.05	0.0069	0.0069	0.0097	0.0097	0.0042	0.0028	0.0042
		0.01	0.0167	0.0139	0.0194	0.0194	0.0167	0.0125	0.0125
		0.005	0.0278	0.0236	0.0250	0.0250	0.0236	0.0181	0.0222
		0.001	0.0500	0.0486	0.0681	0.0681	0.0431	0.0472	0.0569
	1 – AUC		0.0013	0.0011	0.0014	0.0014	0.0009	0.0009	0.0011
f_2 (cubic)	FNR at $\alpha =$	0.05	0.0708	0.0583	0.0750	0.0750	0.0667	0.0514	0.0569
		0.01	0.1514	0.1222	0.1403	0.1403	0.1333	0.1181	0.1306
		0.005	0.1875	0.1458	0.1708	0.1708	0.1639	0.1486	0.1653
		0.001	0.2722	0.2167	0.2694	0.2708	0.2403	0.2097	0.2319
	1 – AUC		0.0177	0.0148	0.0193	0.0193	0.0169	0.0140	0.0158
f_3 (w. flat part)	FNR at $\alpha =$	0.05	0.0528	0.0472	0.0625	0.0625	0.0514	0.0472	0.0528
		0.01	0.1222	0.1000	0.1111	0.1125	0.1000	0.0847	0.0986
		0.005	0.1528	0.1222	0.1361	0.1361	0.1278	0.1125	0.1361
		0.001	0.2472	0.1833	0.2417	0.2403	0.2056	0.1931	0.2236
	1 – AUC		0.0121	0.0103	0.0122	0.0122	0.0108	0.0100	0.0123
f_4 (p.w. linear)	FNR at $\alpha =$	0.05	0.0306	0.0181	0.0361	0.0361	0.0278	0.0236	0.0278
		0.01	0.0861	0.0639	0.0694	0.0694	0.0681	0.0556	0.0681
		0.005	0.1125	0.0847	0.1000	0.1000	0.0931	0.0736	0.0972
		0.001	0.2000	0.1417	0.2069	0.2083	0.1694	0.1500	0.1764
	1 – AUC		0.0060	0.0046	0.0066	0.0066	0.0054	0.0047	0.0056

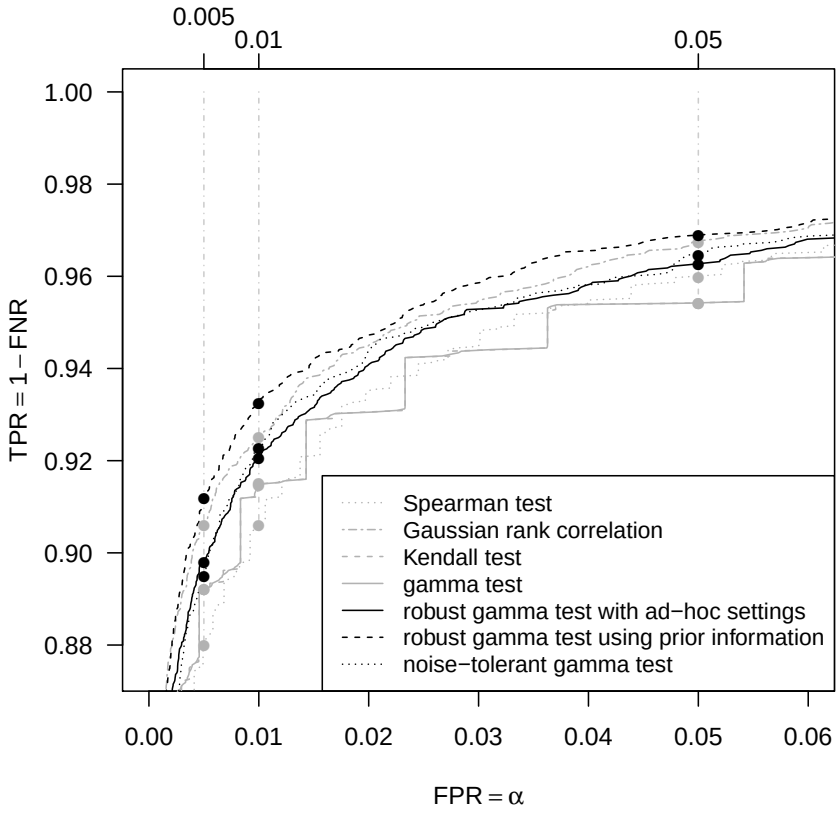

Figure 5: Top-left corner of ROC curves for the seven compared rank correlation tests. The vertical cuts visualize the type II error rates provided by Table 2.

Since the areas under the curve are quite close to 1, Table 2 actually provides $1 - \text{AUC}$ to allow for a better judgment of differences.

Type II error rates as well as the ROC curves only provide reliable information for cases with larger p-values, but they do not allow for any insights which methods give better results in the presence of rather strong correlations which result in very low p-values. Figure 6 provides pairwise scatter plots of p-values and log-transformed p-values obtained for the 2,880 simulated data sets of our comparative study. The scatter plots of plain p-values (boxes above the diagonal) confirm the insights we gained using type II error rates and ROC curves: in all the scatter plots in rows 1, 3, and 4 (corresponding to the Spearman test, the Kendall test, and the classical gamma test) and columns 2 and 5–7 (Gaussian rank correlation test and our three robust gamma variants), the majority of points are

above the diagonal which means that the p-values of the three classical tests are above the p-values of the four robust tests. We also see that the vast majority of the data sets visible in these scatter plots are the small data sets with 10 samples (red dots). All data sets for which small p-values were obtained are concentrated around $(0, 0)$ and cannot be analyzed meaningfully in these plots.

The boxes below the diagonal show scatter plots of log-transformed p-values. In these scatter plots, the differences of high p-values (10^{-2} or higher) are hardly visible, but we get a detailed picture of how the methods compare for data sets with highly significant correlations, therefore, these scatter plots ideally complement the plots of plain p-values. If we compare the log-transformed p-values of the classical gamma test (column 4) with the log-transformed p-values of the three robust gamma variants (rows 5–7), we again see the superiority of the robust variants for small data sets (a certain majority of red dots are below the gray line).

The results in Table 2, the ROC graphs in Figure 5, and the p-value scatter plots in Figure 6 seem to indicate that the robust gamma rank correlation test using prior information outperforms all other methods, while the Gaussian rank correlation test ranks second. To validate these claims statistically, we applied one-sided Wilcoxon tests to all pairs of p-value vectors of the seven tests. The results are summarized in Supplementary Table S2. We clearly see that the robust gamma rank correlation test using prior information gives smaller — more significant — p-values than any test in the comparison (see sixth row of Supplementary Table S2). The Gaussian rank correlation test gives smaller p-values than any other test except the robust gamma rank correlation test using prior information (see second row of Supplementary Table S2). Our family of robust rank correlation tests and the Gaussian rank correlation test give smaller p-values than any of the three classical rank correlation tests.

So we can summarize that the robust gamma rank correlation test using prior information clearly performs best. However, it necessitates knowledge of the underlying noise distribution. If this knowledge is not available, the next best option is the Gaussian rank correlation estimator which outperforms both the robust gamma correlation test with ad-hoc settings and the noise-tolerant gamma rank correlation test. Furthermore, the Gaussian rank correlation test has the advantage that the user need not make any parameter adjustments.

Finally, we have to emphasize again that the error rates, ROC curves and p-values presented in this section were calculated on the basis of our simulated data sets and will consequently be different for other data sets. The specific functions f_1, f_2, f_3 and f_4 that were used to create our data sets have been chosen as representatives for (mostly) monotonic relationships.

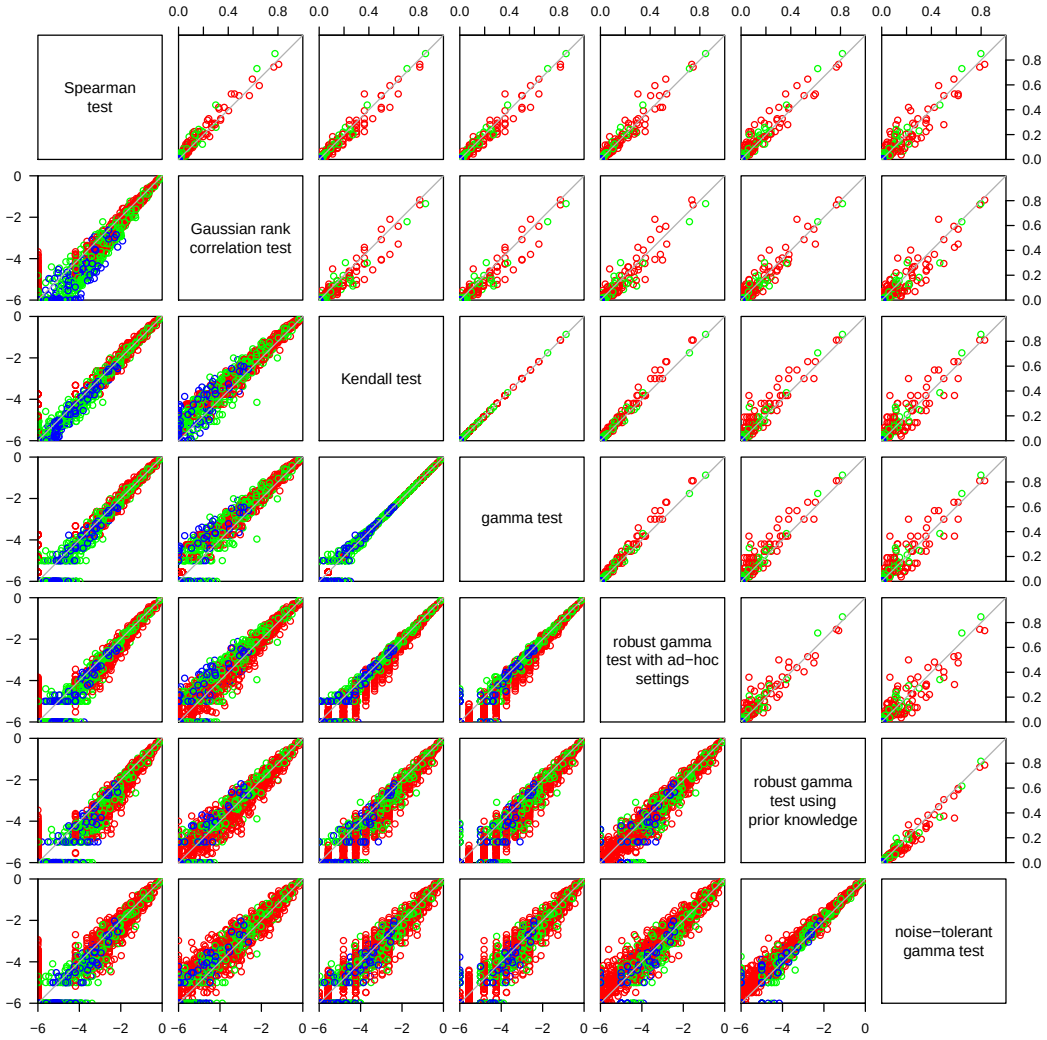

Figure 6: Pairwise comparisons of p-values: each row and each column corresponds to one of the seven tests we compared (see labels along diagonal boxes). The boxes above the diagonal show scatter plots comparing the p-values of two tests on the 2,880 simulated data sets of our comparative study. Each dot in such a plot corresponds to one pair of p-values obtained for one particular data set, where the horizontal component corresponds to the p-value of the "column test" and the vertical component corresponds to the p-value of the "row test". The boxes below the diagonal are to be understood analogously, with the difference that these boxes show scatter plots of log-transformed p-values (to base 10). The colors of the dots encode for the number of samples n of the data sets, where p-value pairs obtained for data sets with 10, 20, and 40 samples are plotted in red, green, and blue, respectively.

7. Experimental Validation of Type I Error Rates

Since the type I error rates are not necessarily equal to the significance threshold α (see discussion in Section 3), we perform another simulation study to estimate type I error rates.

7.1. Data Sets

We created a total number of 6,000 data sets. For three sample sizes, $n = 10$, $n = 20$, and $n = 30$, we created 2,000 paired samples $(x_i, y_i)_{i=1}^n$, where the two vectors $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ were sampled independently from one of the following distributions:

1. Uniform distribution over $[0, 1]$;
2. Normal distribution with $\mu = 0$ and $\sigma = 1$;
3. Mixture of two Gaussians: k samples from a normal distribution $\mathcal{N}(\mu_1, \sigma_1)$, where k is chosen randomly from $\{2, \ldots, n-1\}$, μ_1 is chosen uniformly from $[-1, 0]$ and σ_1 is chosen uniformly from $[0.1, 1]$; the complementary $n - k$ samples are sampled from a normal distribution $\mathcal{N}(\mu_2, \sigma_2)$, where μ_2 is chosen uniformly from $[0, 1]$ and σ_2 is chosen uniformly from $[0.1, 1]$.

Which of the three distribution is used, is decided randomly and independently with equal probability for each individual vector of samples.. Each time a vector is sampled as a mixture of Gaussians, a new random choice of the parameters k, μ_1, σ_1, μ_2, and σ_2 is made.

7.2. Compared Rank Correlation Tests

We applied eight rank correlation tests to all 6,000 data sets:

1. Spearman test
2. Gaussian rank correlation test
3. Kendall test
4. Gamma test
5. Robust gamma rank correlation test with $\bar{T} = T_{\mathbf{M}}$, $R_X = R^{\text{lin}}_{r_x}$, and $R_Y = R^{\text{lin}}_{r_y}$, where the tolerance radii r_x and r_y were chosen as 20% of the interquartile ranges of $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$, respectively.
6. Robust gamma rank correlation test with $\bar{T} = T_{\mathbf{P}}$, $R_X = R^{\text{exp}}_{b_x}$ and $R_Y = R^{\text{exp}}_{b_y}$, where the parameters b_x and b_y were chosen as 20% of the interquartile ranges of $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$, respectively.

7. Robust gamma rank correlation test with $\bar{T} = T_{\mathbf{P}}$, $R_X = R^{\text{Gauss}}_{\sigma_x}$ and $R_Y = R^{\text{Gauss}}_{\sigma_y}$, where the parameters σ_x and σ_y were chosen as 20% of the interquartile ranges of $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$, respectively.
8. Robust gamma rank correlation test with $R_X = R^{\text{crisp}}_{\varepsilon_x}$ and $R_Y = R^{\text{crisp}}_{\varepsilon_y}$, where the tolerance radii ε_x and ε_y were chosen as 20% of the interquartile ranges of $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$, respectively.

Analogously to the experimental evaluation of type II error rates, we performed tests for positive association. Again, all analyses were done using the statistical computing platform R. For the Spearman test and the Kendall test, the standard R function `cor.test()` was used. The classical gamma test, the three robust rank correlation tests, and the Gaussian rank correlation test were performed using our R package `rococo` (see Section 4). For data sets with 10 samples, all permutations were considered. For data sets with 20 or 30 samples, approximate tests were performed using 100,000 random permutations (see Section 3).

7.3. Evaluation Results

Table 3 provides type I error rates estimated from the 6,000 sample data sets. We see that the type I error rates of all eight tests sometimes exceed the significance threshold α. This is particularly true for data sets whose $\mathbf{x}$ and $\mathbf{y}$ vectors have both been sampled as mixtures of Gaussians: all tests have a much higher type I error rate for these data sets (with the obvious effect that the overall type I error rates averaged over all data sets are exceeding the α levels too). For all other combinations of marginal distributions, the type I error rates are quite close to the desired α levels. It seems that all eight tests seem to perform equally well in terms of type I error rates. To investigate this in more depth, we again performed statistical tests evaluating the distributions of p-values. This time we used an unpaired two-sample Wilcoxon test (equivalent to the Mann-Whitney test), since it is mainly the overall distribution of p-values under the null hypothesis that matters for type I error rates. Supplementary Table S3 provides the results in detail. We see that all tests, except the Kendall test, do not show any significant difference in the overall distribution of p-values for the data sets created under the null hypothesis. The Kendall test seems to produce slightly larger p-values than all other methods (see third column of Supplementary Table S3). Thereby, the Kendall test has slightly lower type I error rates. However, the differences are by far not as significant as the differences observed for type II error rates above (see Section 6). So we can safely state that the claims of Section 6 hold true: better type II error rates in the comparison above are not the effect of worse type I error rates.

Table 3: Type I error rates (false positive rates) of the eight compared tests with respect to four different significance levels. The first block gives cumulated rates over all 6,000 data sets. The other six blocks correspond to rates computed for sub-selections of data sets with specific combinations of marginal distributions. As an example, the block labeled "mixt/uni" gives rates for all data sets for which one of the two vectors **x** and **y** was sampled as a mixture of Gaussians, while the other was sampled from a uniform distribution.

			Spearman	Gaussian	Kendall	gamma	robust gamma (linear)	robust gamma (Laplace)	robust gamma (Gauss)	robust gamma (noise-tolerant)
all	FPR at α =	0.05	0.0642	0.0642	0.0577	0.0598	0.0652	0.0653	0.0658	0.0625
		0.01	0.0175	0.0173	0.0167	0.0167	0.0155	0.0163	0.0165	0.0162
		0.005	0.0092	0.0095	0.0083	0.0083	0.0092	0.0088	0.0087	0.0090
		0.001	0.0030	0.0027	0.0022	0.0013	0.0025	0.0028	0.0028	0.0028
norm/norm	FPR at α =	0.05	0.0484	0.0499	0.0456	0.0456	0.0484	0.0484	0.0513	0.0442
		0.01	0.0128	0.0142	0.0142	0.0142	0.0142	0.0100	0.0128	0.0114
		0.005	0.0043	0.0043	0.0085	0.0085	0.0071	0.0071	0.0071	0.0071
		0.001	0.0043	0.0043	0.0028	0.0000	0.0014	0.0028	0.0028	0.0028
norm/mixt	FPR at α =	0.05	0.0517	0.0546	0.0451	0.0473	0.0502	0.0480	0.0495	0.0480
		0.01	0.0087	0.0131	0.0102	0.0102	0.0080	0.0080	0.0087	0.0095
		0.005	0.0058	0.0073	0.0058	0.0058	0.0066	0.0058	0.0058	0.0058
		0.001	0.0015	0.0015	0.0000	0.0000	0.0022	0.0022	0.0022	0.0015
norm/uni	FPR at α =	0.05	0.0556	0.0601	0.0473	0.0495	0.0571	0.0586	0.0586	0.0571
		0.01	0.0135	0.0113	0.0120	0.0120	0.0120	0.0128	0.0128	0.0128
		0.005	0.0075	0.0068	0.0060	0.0060	0.0068	0.0068	0.0068	0.0060
		0.001	0.0023	0.0015	0.0015	0.0008	0.0008	0.0008	0.0008	0.0008
mixt/mixt	FPR at α =	0.05	0.1732	0.1627	0.1566	0.1611	0.1777	0.1762	0.1792	0.1627
		0.01	0.0633	0.0572	0.0542	0.0542	0.0557	0.0617	0.0602	0.0587
		0.005	0.0422	0.0437	0.0331	0.0331	0.0407	0.0377	0.0377	0.0392
		0.001	0.0151	0.0120	0.0136	0.0105	0.0136	0.0166	0.0166	0.0166
mixt/uni	FPR at α =	0.05	0.0472	0.0457	0.0457	0.0480	0.0511	0.0526	0.0519	0.0511
		0.01	0.0132	0.0132	0.0124	0.0124	0.0093	0.0124	0.0108	0.0101
		0.005	0.0031	0.0039	0.0023	0.0023	0.0023	0.0023	0.0023	0.0031
		0.001	0.0000	0.0008	0.0000	0.0000	0.0008	0.0000	0.0000	0.0000
uni/uni	FPR at α =	0.05	0.0472	0.0440	0.0409	0.0425	0.0440	0.0456	0.0425	0.0440
		0.01	0.0110	0.0094	0.0126	0.0126	0.0110	0.0094	0.0110	0.0110
		0.005	0.0031	0.0016	0.0047	0.0047	0.0031	0.0047	0.0031	0.0047
		0.001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0016

8. Concluding Remarks

This paper introduced a family of rank correlation tests based on our robust gamma rank correlation coefficients [5, 6]. A detailed comparative study with a large number of simulated data sets was carried out to validate if and under which particular circumstances the new tests and a test based on the Gaussian rank correlation estimator outperform the established rank correlation tests, such as the Spearman test, the Kendall test, and the classical gamma test. We have seen that a clear improvement of type II error rates is achieved. This is mainly due to the fact that the robust tests are less sensitive to noise, especially for small numbers of samples. At the same time, the type I error rates of the robust gamma rank correlation tests, the Gaussian rank correlation test, and the classical rank correlation tests are in comparable ranges.

We are convinced that the empirical arguments raised in this paper provide sufficient practical justification in favor of robust rank correlation tests. The remaining challenges are more on the theoretical side. One important challenge is to achieve a better understanding of the null distributions, although we are not sure that substantial progress can be made in this direction. The study of mathematical properties of the family of robust rank correlation coefficients is for sure not complete either. A recent paper [33] has made some progress, but some more issues remain open, e.g. which of the properties discussed in [14] are fulfilled by robust rank correlation coefficients. Another question of practical importance is whether correlation matrices in a multi-variate settings are positive semi-definite or not.

Finally, we note that although this paper deals completely with traditional statistics, it demonstrates that methods of traditional statistics can be improved by the integration of concepts from fuzzy set theory (in line with [11]), in our case, by replacing crisp relations by fuzzy/graded ones.

Acknowledgements

The authors gratefully acknowledge partial support of *COST Action IC0702 "SoftStat — Combining Soft Computing Techniques and Statistical Methods to Improve Data Analysis Solutions"*. Furthermore, the authors thank the anonymous reviewers for providing highly valuable suggestions that helped to improve the manuscript.

References

[1] H. Abdi, Coefficients of correlation, alienation and determination, in: N.J. Salkind (Ed.), Encyclopedia of Measurement and Statistics, Sage, Thousand Oaks, CA, 2007.

[2] H. Abdi, The Kendall rank correlation coefficient, in: N.J. Salkind (Ed.), Encyclopedia of Measurement and Statistics, Sage, Thousand Oaks, CA, 2007.

[3] U. Bodenhofer, A similarity-based generalization of fuzzy orderings preserving the classical axioms, Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 8 (2000) 593–610.

[4] U. Bodenhofer, M. Demirci, Strict fuzzy orderings with a given context of similarity, Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 16 (2008) 147–178.

[5] U. Bodenhofer, F. Klawonn, Towards robust rank correlation measures for numerical observations on the basis of fuzzy orderings, in: M. Štěpnička, V. Novák, U. Bodenhofer (Eds.), Proc. 5th Conference of the European Society for Fuzzy Logic and Technology, volume I, Ostrava, pp. 321–327.

[6] U. Bodenhofer, F. Klawonn, Robust rank correlation coefficients on the basis of fuzzy orderings: initial steps, Mathware Soft Comput. 15 (2008) 5–20.

[7] U. Bodenhofer, M. Krone, RoCoCo: an R package implementing a robust rank correlation coefficient and a corresponding test, 2011. Software available at `http://www.bioinf.jku.at/software/rococo/`.

[8] K. Boudt, J. Cornelissen, C. Croux, The Gaussian rank correlation estimator: robustness properties, Stat. Comput. 22 (2012) 471–483.

[9] A.W. Bowman, M.C. Jones, I. Gijbels, Testing monotonocity of regression, J. Comput. Graph. Stat. 7 (1998) 489–500.

[10] J.J. Buckley, Fuzzy statistics: hypothesis testing, Soft Computing 9 (2005) 512–518.

[11] R. Coppi, M.A. Gil, H.A.L. Kiers, The fuzzy approach to statistical analysis, Comput. Stat. Data Anal. 51 (2006) 1–14.

[12] P. Davies, U. Gather, Breakdown and groups (with discussion), Ann. Stat. 33 (2005) 977–1035.

[13] N. de Klerk, Commentary: Spearman's 'The proof and measurement of association between two things, Int. J. Epidemiol. 39 (2010) 1159–1161.

[14] C. Elzinga, H. Wang, Z. Lin, Y. Kumar, Concordance and consensus, Inform. Sci. 181 (2011) 2529–2549.

[15] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (2006) 861–874.

[16] P. Filzmoser, R. Viertl, Testing hypotheses with fuzzy data: the fuzzy p-value, Metrika 59 (2004) 21–29.

[17] C.J. Geyer, G.D. Meeden, Fuzzy and randomized confidence intervals and p-values, Stat. Sci. 20 (2005) 358–366.

[18] S. Ghosal, A. Sen, A.W. van der Vaart, Testing monotonocity of regression, Ann. Stat. 28 (2000) 1054–1082.

[19] L.A. Goodman, W.H. Kruskal, Measures of association for cross classifications, J. Amer. Statist. Assoc. 49 (1954) 732–764.

[20] P. Hall, N.E. Heckman, Testing for monotonocity of a regression mean by calibrating linear functions, Ann. Stat. 28 (2000) 20–39.

[21] D.C. Hoaglin, F. Mosteller, J.W. Tukey, Understanding Robust and Exploratory Data Analysis, John Wiley & Sons, New York, 2000.

[22] M.Z. Huang, H.B. Li, X.M. Nie, C.M. Jiang, H. Ming, D.Q. Li, X.Y. Wu, Analysis of the dose-response relationship between high-risk human papillomavirus viral load and cervical lesions, Trans. R. Soc. Trop. Med. Hyg. 103 (2009) 779–784.

[23] P.J. Huber, E.M. Ronchetti, Robust Statistics, Wiley Series in Probability and Statistics, John Wiley & Sons, New York, second edition, 2009.

[24] G. Imre, D.S. Fokkema, J.A. Den Boer, G.J. Ter Horst, Dose-response characteristics of ketamine effect on locomotion, cognitive function and central neuronal activity, Brain. Res. Bull. 69 (2006) 338–345.

[25] M.G. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81–93.

[26] M.G. Kendall, Rank Correlation Methods, Charles Griffin & Co., London, third edition, 1962.

[27] E.P. Klement, R. Mesiar, E. Pap, Triangular Norms, volume 8 of *Trends in Logic*, Kluwer Academic Publishers, Dordrecht, 2000.

[28] H.W. Koh, E. Hüllermeier, Mining gradual dependencies based on fuzzy rank correlation, in: C. Borgelt, W. Trutschnig, G. Gonzalez-Rodriguez, M.A. Lubiano, M.A. Gil, P. Grzegorzewski, O. Hryniewicz (Eds.), Combining Soft Computing and Statistical Methods in Data Analysis, volume 77 of *Advances in Intelligent and Soft Computing*, Springer, Dordrecht, 2010, pp. 379–386.

[29] W.H. Kruskal, Ordinal measures of association, J. Amer. Statist. Assoc. 53 (1958) 814–861.

[30] R.A. Maronna, D.R. Martin, V.J. Yohai, Robust Statistics: Theory and Methods, Wiley Series in Probability and Statistics, John Wiley & Sons, Toronto, ON, 2006.

[31] M. Montenegro, M.R. Casals, M.A. Lubiano, M.A. Gil, Two-sample hypothesis tests of means of a fuzzy random variable, Inform. Sci. 133 (2001) 89–100.

[32] K. Pearson, Notes on the history of correlation, Biometrika 13 (1920) 25–45.

[33] M.D. Ruiz, E. Hüllermeier, A formal and empirical analysis of the fuzzy gamma rank correlation coefficient, Inform. Sci. 206 (2012) 1–17.

[34] C. Spearman, The proof and measurement of association between two things, Am. J. Psychol. 15 (1904) 72–101.

[35] C. Spearman, Demonstration of formulae for true measurement of correlation, Am. J. Psychol. 18 (1907) 161–169.

[36] D. Steinhauser, B.H. Junker, A. Luedemann, J. Selbig, J. Kopka, Hypothesis-driven approach to predict transcriptional units from gene expression data, Bioinformatics 20 (2004) 1928–1939.

[37] E.A. Thompson, C.J. Geyer, Fuzzy p-values in latent variable problems, Biometrika 94 (2007) 49–60.