International Journal of Approximate Reasoning 56 (2015) 108–121

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Generalizing the Wilcoxon rank-sum test for interval data

Julien Perolat [a], Inés Couso [b], Kevin Loquin [a], Olivier Strauss [a,*]

[a] LIRMM Université Montpellier II, France
[b] Universidad de Oviedo, Statistics Department, Spain

ARTICLE INFO

Article history:
Received 7 November 2013
Received in revised form 31 July 2014
Accepted 2 August 2014
Available online 7 August 2014

Keywords:
Statistical hypothesis test
Interval-valued imprecise data
Bipolar decision

ABSTRACT

Here we propose an adaption of Wilcoxon's two-sample rank-sum test to interval data. This adaption is interval-valued: it computes the minimum and maximum values of the statistic when we rank the set of all feasible samples (all joint samples compatible with the initial set-valued information). We prove that these bounds can be explicitly computed using a very low computational cost algorithm. Interpreting this generalized test is straightforward: if the obtained interval-valued p-value is on one side of the significance level, we will be able to make a decision (reject/no reject). Otherwise, we will conclude that our information is too vague to lead to a clear decision.

Our method is also applicable to quantized data: in the presence of quantized information, the joint sample may contain a high proportion of draws, which can prevent the test from drawing a clear conclusion. According to the usual convention, when there are ties, the ranks for the observations in a tie are taken to be the average of the ranks for those observations. This convention can lead to wrong conclusions. Here, we consider the family of all possible rank permutations, such that a sample containing ties will not just be associated with a single value, but rather with a collection of values for the Wilcoxon's rank-sum statistic, with each one of them being associated with a different p-value. When the impact of quantization is too high to lead to a clear decision, our test provides an interval-valued p-value that includes the chosen significance level. It indicates that there is no clear conclusion according to this test.

Two different experiments exemplify the properties of the generalized test: the first one illustrates its ability to avoid wrong decisions in the presence of quantized data. The second one shows the performance of the generalized test when used with interval data.

1. Introduction

The Wilcoxon rank-sum test [17], also know as Mann–Whitney U test, is a non-parametric hypothesis test used to check whether or not two independent samples containing n and m elements correspond to the same distribution. It does not require the data normality assumption. It can thus be regarded as an alternative to the two-sample t-test when the normality of the data clearly cannot be assumed. The Wilcoxon rank-sum statistic W is computed as follows: the $n+m$ observations of the two independent samples are combined in a single dataset. The elements of this dataset are sorted from

* Corresponding author.
E-mail addresses: Julien.Perolat@supelec.fr (J. Perolat), couso@uniovi.es (I. Couso), Kevin.Loquint@lirmm.fr (K. Loquin), Olivier.Strauss@lirmm.fr (O. Strauss).

http://dx.doi.org/10.1016/j.ijar.2014.08.001
0888-613X/

smallest to largest. If there are ties, i.e. duplicated values in the combined dataset, the ranks for the observations in a tie are taken to be the average of the ranks for those observations.

The Wilcoxon statistic, W, is calculated as the sum of the ranks for the n observations from the first population. If the null hypothesis of identical population distributions is true, n ranks from the first population are just a random sample from the $n+m$ integers $1,\ldots,n+m$. Under this null hypothesis, the expectation and variance of W are, respectively:

$$\mu_0 = \frac{n(n+m+1)}{2} \quad \text{and} \quad \sigma_0^2 = \frac{nm(n+m+1)}{12}.$$

Furthermore, when both sample sizes are sufficiently large ($n > 10$ and $m > 10$, by convention), the distribution of the statistic $T = \frac{W-\mu_0}{\sigma_0}$ is assumed to be Gaussian (with null expectation, and variance equal to 1). Intuitively, if W is much smaller (or larger) than μ_0, this is evidence that the null hypothesis is false and, in fact, that the considered samples come from distinct populations. Under this assumption, the critical value (or p-value) is calculated according to the following equality:

$$p(w) = 2\left[1 - \phi\left(\frac{|w-\mu_0|}{\sigma_0}\right)\right],$$

where w denotes the value of the rank-sum statistic in the sample, and ϕ denotes the cumulative distribution function of the standard normal. For a specific significance level $\alpha \in (0,1)$, the null hypothesis will be rejected whenever $p(w) < \alpha$.

In this paper, we explore an alternative to the convention that assigns an average rank to tied values. If there is a high proportion of draws in the sample (which can easily occur in the presence of discrete or quantized data), considering a different (but also compatible) sequence of rank assignments can lead to a completely different final decision. Indeed, the computed rank-sum statistic w, and therefore the associated p-value, $p(w)$, may vary significantly. In this paper, we consider all the possible rank assignments, so that a sample containing ties, will not just be associated with a single value (computed by averaging the ranks of the tied elements), but rather to a collection of values for the Wilcoxon rank sum statistic, with each one of them being associated with a different p-value. The decision process thus becomes more expressive. If all of the computed p-values are below a fixed significance level α, we will reject the null hypothesis. Likewise, if the p-values are all above α, then one would conclude that there is little evidence for the alternative hypothesis to be true (we "accept" the null hypothesis). There is a now a third possible answer: if the set of p-values and the significance level do overlap, we cannot make a decision because our information on the original values that led to that "quantized" data is not sufficiently accurate. In other words, quantization hides the information and thus the problem is no longer decidable.

The procedure described in the previous paragraph can be applied, more generally, to situations where, instead of point-valued sample observations, we are provided with interval-valued sample observations, with each one containing the true instance.

Interval-valued observations can account for a known defect in the measurement system (see e.g. [16,18,14] and references therein), quantization of data, guaranteed estimation ([9,8,10]) or for imprecision in the model (see e.g. [15,13]). We aim to cover the whole collection of samples compatible with our incomplete information in order to calculate the set of feasible rank-sum statistics. A p-value corresponds to each of those rank-sum statistics. Thus, our information about the critical value associated with the true (imprecisely observed) joint sample will be determined by a subset of $[0,1]$, where it certainly belongs. When the initial information is not very accurate, the set of possible p-values will be larger. The decision procedure is similar to the previously presented approach. When the whole set is on one side of the $\alpha-$level, we will be able to make a decision, otherwise we will not. This approach was recently considered by several authors (see, for instance [1–5]). Specifically, Denoeux et al. [3] applied this procedure to the Wilcoxon rank-sum test, in the presence of fuzzy data. This encompasses, as a particular case, the case of interval-valued data. The same idea of looking for the least and most favorable cases within the imprecise dataset can be also found in the recent literature (see [1,6,7,12,11] among others).

One of the main pitfalls in this framework is to compute the bounds of the set of feasible p-values. This task can be relatively computationally expensive, depending on the test we are generalizing. As Denoeux et al. [3] state, a simplistic approach used to solve the problem of determining the maximum and minimum values for the statistic and its corresponding p-value, in this kind of rank-based tests, might be to generate all rank assignments that are compatible with the available incomplete information. However, this approach is intractable, as they point out, due to the potentially exponential number of different rankings, which can reach $n!$ in the empty information limit case. Thus, they propose to use a Monte-Carlo simulation method to approximate the bounds of the Wilcoxon rank-sum statistic. Such a method would require many iterations for large n and m values, but would not provide exact values of the bounds of the set of feasible values for the statistic. In this paper, we try to overcome this issue. We present a procedure to calculate exact values for those bounds. Furthermore, we show that the complexity of such a procedure is comparable to the calculation of the Wilcoxon rank-sum statistic for single-valued data, because each of the bounds just involves calculation of the Wilcoxon rank-sum test for a specific extreme sample. We provide two explicit and computationally simple algorithms to compute those bounds.

We propose two experiments to illustrate the properties of this generalization. The first experiment highlights the ability of the imprecise test to discard wrong decisions due to quantization in precise data. The second experiment shows the behavior of the test when interval-valued data are considered.

2. Generalization of the Wilcoxon rank-sum test

In this section, we describe the procedure to make decisions on the basis of a Wilcoxon rank-sum test when provided with an imprecise sample expressed in terms of interval data. Subsection 2.1 describes an easy method to determine the minimum and maximum of the set of feasible values for the rank-sum statistic. Subsection 2.2 deals with calculation of bounds for p-values while providing a procedure to make decisions based upon this information.

Let us first outline the nomenclature we will use in the section. $\mathbf{X} = X_1 \times \cdots \times X_i \times \cdots \times X_n$ and $\mathbf{Y} = Y_1 \times \cdots \times Y_j \times \cdots \times Y_m$ denote our incomplete information about ill-known samples $\mathbf{x}^0 = (x_1^0 \dots x_i^0 \dots x_n^0) \in \mathbb{R}^n$ and $\mathbf{y}^0 = (y_1^0 \dots y_j^0 \dots y_m^0)$, where each X_i and each Y_j is a closed interval of the form $X_i = [\underline{x}_i, \bar{x}_i]$ and $Y_j = [\underline{y}_j, \bar{y}_j]$ representing incomplete information about the value of each component. $\underline{\mathbf{x}} = (\underline{x}_1 \dots \underline{x}_i \dots \underline{x}_n)$ and $\bar{\mathbf{x}} = (\bar{x}_1 \dots \bar{x}_i \dots \bar{x}_n)$ denotes the pair of samples included in $\mathbf{X}$ that are formed by the minimum (resp. the maximum) of the interval data. Analogously, we use the notation $\underline{\mathbf{y}} = (\underline{y}_1 \dots \underline{y}_j \dots \underline{y}_m)$ and $\bar{\mathbf{y}} = (\bar{y}_1 \dots \bar{y}_j \dots \bar{y}_m)$. Let $\mathbf{xy}$ and $\mathbf{XY}$ denote the joint samples associated with samples $\mathbf{x}$ and $\mathbf{y}$ as well as samples $\mathbf{X}$ and $\mathbf{Y}$, respectively. $\mathbf{x} \in \mathbf{X}$ represents the fact that for any $i \in \{1, \dots, n\}$, $x_i \in X_i$ (the same for $\mathbf{y} \in \mathbf{Y}$ and $\mathbf{xy} \in \mathbf{XY}$).

2.1. Calculation of the bounds of the rank-sum statistic

Consider an arbitrary pair of vectors $\mathbf{x} = (x_1, \dots, x_n) \in \mathbf{X}$ and $\mathbf{y} = (y_1, \dots, y_m) \in \mathbf{Y}$. For every $v \in \{x_1, \dots, x_n, y_1, \dots, y_m\}$, $\underline{r}_{\mathbf{xy}}(v)$ and $\bar{r}_{\mathbf{xy}}(v)$ respectively denote the minimum and maximum possible rank values of $v \in \mathbb{R}$ in the joint sample $\mathbf{xy}$, i.e.

$$\underline{r}_{\mathbf{xy}}(v) = \#\{k \in \{1, \dots, n\} : x_k < v\} + \#\{l \in \{1, \dots, n\} : y_l < v\} + 1,$$

and

$$\bar{r}_{\mathbf{xy}}(v) = \#\{k \in \{1, \dots, n\} : x_k \le v\} + \#\{l \in \{1, \dots, n\} : y_l \le v\}.$$

$R_{\mathbf{xy}}(v)$ denotes the (finite) set of rank values associated with v, i.e.:

$$R_{\mathbf{xy}}(v) = \{n \in \mathbb{N} : \underline{r}_{\mathbf{xy}}(v) \le n \le \bar{r}_{\mathbf{xy}}(v)\} \subseteq \{1, \dots, n+m\}.$$

Any bijective mapping $s_{\mathbf{xy}} : \{1, \dots, n+m\} \to \{1, \dots, n+m\}$ (permutation of $n+m$ numbers) satisfying the constraints $s_{\mathbf{xy}}(i) \in R_{\mathbf{xy}}(x_i)$, $i = 1, \dots, n$, $s_{\mathbf{xy}}(j+n) \in R_{\mathbf{xy}}(y_j)$, $j = 1, \dots, m$, is referred to as a *rank assignment*.

Readers should note that, for every pair $v, v' \in \{x_1, \dots, x_n, y_1, \dots, y_m\}$ with $v \neq v'$, we have $R_{\mathbf{xy}}(v) \cap R_{\mathbf{xy}}(v') = \emptyset$. Moreover, for an arbitrary v, the cardinality of $R_{\mathbf{xy}}(v)$ is equal to the number of elements in the joint sample taking the value v, i.e.:

$$\#R_{\mathbf{xy}}(v) = \#\{i \in \{1, \dots, n\} : x_i = v\} + \#\{j \in \{1, \dots, m\} : y_j = v\} = \#I(v), \tag{1}$$

where $I(v) = \{k \in \{1, \dots, n+m\} : x_k = v \text{ or } y_{k-n} = v\}$ denotes the collection of indices associated with v. Thus, the families of sets $\{R_{\mathbf{xy}}(v) : v \in \{x_1, \dots, x_n, y_1, \dots, y_m\}\}$ and $\{I(v) : v \in \{x_1, \dots, x_n, y_1, \dots, y_m\}\}$ constitute two (usually different) partitions of the set of indices $\{1, \dots, n+m\}$. We can easily check that there is at least one rank assignment $s_{\mathbf{xy}}$ satisfying the above constraints. For each $v \in \{x_1, \dots, x_n, y_1, \dots, y_m\}$, it assigns a rank value included in $R_{\mathbf{xy}}(v)$ to every element in $I(v)$. $\Sigma_{\mathbf{xy}}$ denotes the family of all rank assignments associated with the sample $\mathbf{xy}$. Of course, if there are no ties, it will be a singleton. The mapping $w : \bigcup_{\mathbf{xy} \in \mathbf{XY}} \Sigma_{\mathbf{xy}} \to \mathbb{N}$ will assign, to each specific rank assignment, the sum of ranks of the first n elements, $w(s_{\mathbf{xy}}) = \sum_{i=1}^{n} s_{\mathbf{xy}}(i)$. Finally, $W_{\mathbf{xy}}$ denotes the family of all possible values for the Wilcoxon rank-sum statistic associated to the sample $\mathbf{xy}$. Formally:

$$W_{\mathbf{xy}} = \{w(s_{\mathbf{xy}}) : s_{\mathbf{xy}} \in \Sigma_{\mathbf{xy}}\}.$$

For an arbitrary sample $\mathbf{xy} \in \mathbf{XY}$, we can easily find the minimum of $W_{\mathbf{xy}}$ if we consider any permutation assigning the least positions in $R_{\mathbf{xy}}(v)$ to the indices in $I(v)$ associated with those instances coming from the first sample, $\{i \in \{1, \dots, n\} : x_i = v\} \subseteq I(v)$, and the largest ones to those elements coming from the second one. Formally, such a permutation will satisfy the following conditions:

$$s_{\mathbf{xy}}(i) < s_{\mathbf{xy}}(j), \quad \text{if } i \le n < j, \text{ and } x_i = y_{j-n}.$$

Analogously, we can easily find the maximum of $W_{\mathbf{xy}}$ by considering any rank assignment satisfying the constraints:

$$s_{\mathbf{xy}}(i) > s_{\mathbf{xy}}(j), \quad \text{if } i \le n < j, \text{ and } x_i = y_{j-n}.$$

Next, we prove that, for an arbitrary pair of samples $\mathbf{x} \in \mathbf{X}$, $\mathbf{y} \in \mathbf{Y}$, the set of Wilcoxon values, $W_{\mathbf{xy}}$, assigned to the joint sample $\mathbf{xy}$ is bounded from below by the minimum of the set of values associated with the sample $\underline{\mathbf{x}}\bar{\mathbf{y}}$ and bounded from above by the maximum of the set of values associated with $\bar{\mathbf{x}}\underline{\mathbf{y}}$. Formally, we state this as follows:

Theorem 2.1. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be two imprecise samples of size n and m, respectively. For any pair of precise samples* $\mathbf{x} \in \mathbf{X}$ *and* $\mathbf{y} \in \mathbf{Y}$*, the following inequalities hold:*

$$\min W_{\underline{\mathbf{x}}\bar{\mathbf{y}}} \leq \min W_{\mathbf{xy}} \leq \max W_{\mathbf{xy}} \leq \max W_{\bar{\mathbf{x}}\underline{\mathbf{y}}}.$$

Proof. It suffices to check that, for any $w \in W_{\mathbf{xy}}$ we have $w'_1 \in W_{\underline{\mathbf{x}}\mathbf{y}}$ and $w''_1 \in W_{\bar{\mathbf{x}}\mathbf{y}}$ such that $w'_1 \leq w \leq w''_1$ and $w'_2 \in W_{\mathbf{x}\bar{\mathbf{y}}}$ and $w''_2 \in W_{\mathbf{x}\underline{\mathbf{y}}}$ such that $w'_2 \leq w \leq w''_2$. We just prove the cases of w'_1 and w'_2 to deduce the left inequality of Theorem 2.1. The demonstration for the right inequality is similar. The central inequality is obvious.

Let us consider an arbitrary sample $\mathbf{xy} \in \mathbf{XY}$, and a specific rank assignment $s_{\mathbf{xy}}$, and let $w = w(s_{\mathbf{xy}})$. We just need to check that, if we consider a new joint sample $\mathbf{x}'\mathbf{y}$, where $\mathbf{x}'$ is defined as follows: $x'_k = \underline{x}_k$ and $x'_j = x_j$, $\forall j \neq k$ then there is a rank assignment $s_{\mathbf{x}'\mathbf{y}}$ such that $w(s_{\mathbf{x}'\mathbf{y}}) \leq w(s_{\mathbf{xy}})$ and something similar happens if, for an arbitrary $l \in \{1, \ldots, m\}$ we replace y_l by $y'_l = \bar{y}_l$.

Let us prove the first part. We assign to $x'_k = \underline{x}_k$ the rank value $r = s_{\mathbf{x}'\mathbf{y}}(k) := \underline{r}_{\mathbf{x}'\mathbf{y}}(x_k) = \#\{i \in \{1, \ldots, n\} : x'_i < x'_k\} + \#\{r \in \{1, \ldots, m\} : y_r < x'_k\} + 1$ (note that, according to this procedure, if $x'_k = \underline{x}_k$ coincides with some other element in the joint sample $\mathbf{x}'\mathbf{y}$, we are assigning it the smallest possible rank value). We easily observe that $s_{\mathbf{x}'\mathbf{y}}(k) = s_{\mathbf{xy}}(k) - s$ with $s \geq 0$ (if $x'_k = \underline{x}_k$ does coincide with some other element in the joint sample $\mathbf{x}'\mathbf{y}$). Furthermore, for those indices $i \in \{1, \ldots, n\}$ such that $s_{\mathbf{xy}}(i) \in \{r, \ldots, r + s - 1\}$, we have $s_{\mathbf{x}'\mathbf{y}}(i) = s_{\mathbf{xy}}(i) + 1$. There are at most s of those indices. Furthermore, for the rest of the indices i, we have $s_{\mathbf{x}'\mathbf{y}}(i) = s_{\mathbf{xy}}(i)$. We therefore deduce that $w(s_{\mathbf{x}'\mathbf{y}}) \leq w(s_{\mathbf{xy}})$.

The proof of the second part is a bit shorter because, when we replace y_l by $y'_l = \bar{y}_l$, none of the rank values, $s_{\mathbf{xy}'}(i)$ are higher than the corresponding initial one, $s_{\mathbf{xy}}(i)$ and therefore, $w(s_{\mathbf{xy}'}) \leq w(s_{\mathbf{xy}})$. Regarding potential draws between elements in the sample $\mathbf{xy}'$, an analogous procedure applies in order to find the sample with the maximum sum of ranks: we place y'_l in the highest possible position, i.e., $s_{\mathbf{xy}'}(l) = \#\{i \in \{1, \ldots, n\} : x_i \leq y'_l\} + \#\{j \in \{1, \ldots, m\} : y_j \leq y'_l\}$. □

According to Theorem 2.1, for any joint sample $\mathbf{xy}$ compatible with the imprecise information provided by $\mathbf{XY}$, any rank summation assigned to it, $w(s_{\mathbf{xy}}) \in W_{\mathbf{xy}}$ is an integer value bounded by $\min W_{\underline{\mathbf{x}}\bar{\mathbf{y}}}$ and $\max W_{\bar{\mathbf{x}}\underline{\mathbf{y}}}$. Now we will go further and prove that, for any integer r bounded by those two numbers, there is at least one joint sample $\mathbf{xy} \in \mathbf{XY}$ and a rank assignment $s_{\mathbf{xy}} \in \Sigma_{\mathbf{xy}}$ such that $r = w(s_{\mathbf{xy}})$. This result will impact calculation of the set of possible p-values derived from the (imprecise) information provided by $\mathbf{XY}$, as we will show later.

Theorem 2.2. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be two imprecise samples of size n and m. Let* $W_{\mathbf{XY}} = \bigcup_{\mathbf{xy} \in \mathbf{XY}} W_{\mathbf{xy}}$ *denote the set of all possible values for the rank-sum statistic. Then:*

$$W_{\mathbf{XY}} = \{r \in \mathbb{N} : \min W_{\underline{\mathbf{x}}\bar{\mathbf{y}}} \leq r \leq \max W_{\bar{\mathbf{x}}\underline{\mathbf{y}}}\}.$$

Proof. It suffices to check that for every $r \in \mathbb{N}$ satisfying the conditions 1) $\min W_{\underline{\mathbf{x}}\bar{\mathbf{y}}} \leq r < \max W_{\bar{\mathbf{x}}\underline{\mathbf{y}}}$ and 2) $r \in W_{\mathbf{XY}}$, the value $r + 1$ also belongs to $W_{\mathbf{XY}}$. Let us consider an arbitrary r satisfying these restrictions. According to the second condition, there is at least one sample $\mathbf{xy} \in \mathbf{XY}$ and a rank assignment $s_{\mathbf{xy}} \in \Sigma_{\mathbf{xy}}$ such that $r = w(s_{\mathbf{xy}})$. Furthermore, according to the first condition, there is at least one pair $(i, j) \in \{1, \ldots, n\} \times \{1, \ldots, m\}$, and a rank assignment $s_{\bar{\mathbf{x}}\underline{\mathbf{y}}} \in \Sigma_{\bar{\mathbf{x}}\underline{\mathbf{y}}}$ such that $s_{\mathbf{xy}}(i) < s_{\mathbf{xy}}(j + n)$, but $s_{\bar{\mathbf{x}}\underline{\mathbf{y}}}(i) > s_{\bar{\mathbf{x}}\underline{\mathbf{y}}}(j + n)$. We can assume, without loss of generality, that $x_i < y_j$ (otherwise, we can easily select a new rank assignment $s'_{\mathbf{xy}} \in \Sigma_{\mathbf{xy}}$ such that $w(s'_{\mathbf{xy}}) = w(s_{\mathbf{xy}}) + 1$, just exchanging indices between linked values). We will also assume, without loss of generality, that $s_{\mathbf{xy}}(i) = \max_{k : x_k = x_i} s_{\mathbf{xy}}(k)$ and $s_{\mathbf{xy}}(j + n) = \min_{l : y_l = y_j} s_{\mathbf{xy}}(l + n)$. Let us now select two indices $k^* \in \{1, \ldots, n\}$ and $l^* \in \{1, \ldots, m\}$ satisfying the respective conditions $x_{k^*} = \max\{x_k : x_k < y_j\}$ and $y_{l^*} = \min\{y_l : y_l > x_i\}$. We can find at least a pair of indices satisfying those conditions because the above sets of values do respectively contain x_i and y_j. Furthermore, we observe that $x_{k^*} \geq x_i$ and $y_{l^*} \leq y_j$. We will consider three possible cases:

- Case 1, $\bar{x}_i \geq y_{l^*}$. – Let us consider a new sample $\mathbf{x}' \in \mathbf{X}$ defined as $x'_i = y_{l^*}$ and $x'_k = x_k$, $\forall k \neq i$. Let us consider the rank assignment $s_{\mathbf{x}'\mathbf{y}} \in \Sigma_{\mathbf{x}'\mathbf{y}}$ determined as follows:

$$s_{\mathbf{x}'\mathbf{y}}(t) = \begin{cases} s_{\mathbf{xy}}(l^* + n) - 1 & \text{if } t = l^* + n, \\ s_{\mathbf{xy}}(l^* + n) & \text{if } t = i, \\ s_{\mathbf{xy}}(t) - 1, & \text{if } t \in \{1, \ldots, n\} \setminus \{i\} \text{ and } x_i < x_t \leq y^*_l, \\ s_{\mathbf{xy}}(t) & \text{otherwise.} \end{cases}$$

- Case 2, $\bar{x}_i < y_{l^*}$ and $\underline{y}_j \leq x_{k^*}$. – Let us consider a new sample $\mathbf{y}' \in \mathbf{Y}$ defined as $y'_j = x_{k^*}$ and $y'_l = y_l$, $\forall l \neq j$. Let us consider the rank assignment $s_{\mathbf{xy}'} \in \Sigma_{\mathbf{xy}'}$ determined as follows:

$$s_{\mathbf{xy}'}(t) = \begin{cases} s_{\mathbf{xy}}(k^*) + 1 & \text{if } t = k^*, \\ s_{\mathbf{xy}}(k^*) & \text{if } t = j + n, \\ s_{\mathbf{xy}}(t) + 1 & \text{if } x_{k^*} \leq y_{t-n} \leq y_j \\ s_{\mathbf{xy}}(t) & \text{otherwise.} \end{cases}$$

- Case 3, $\bar{x}_i < y_{l^*}$ and $\underline{y}_j > x_{k^*}$. – Let us consider a new joint sample $\mathbf{x}'\mathbf{y}' \in \mathbf{XY}$ defined as $x'_i = \underline{y}_j = y'_j$ and $x'_k = x_k$ $\forall k \neq i$ and $y'_l = y_l$, $\forall l \neq j$. Since we assumed that $s_{\bar{\mathbf{x}}\underline{\mathbf{y}}}(i) > s_{\bar{\mathbf{x}}\underline{\mathbf{y}}}(j+n)$, we deduce that $\bar{x}_i \geq \underline{y}_j$ and therefore $\bar{x}_i \geq x'_i$. Furthermore, $x_{k^*} \geq x_i$ by definition, and we also assume that $\underline{y}_j > x_{k^*}$, so therefore, $x'_i = \underline{y}_j > x_i \geq \underline{x}_i$. We can thus guarantee that x'_i belongs to X_i. Let us consider the following rank assignment $s_{\mathbf{x}'\mathbf{y}'} \in \Sigma_{\mathbf{x}'\mathbf{y}'}$, that assigns the maximum possible rank, $s_{\mathbf{x}'\mathbf{y}'}(i) = \max_{k \in I(\underline{y}_j)} s_{\mathbf{x}'\mathbf{y}'}(k)$, to $x'_i = \underline{y}_j$, and the preceding position $s_{\mathbf{x}'\mathbf{y}'}(j+n) = \max_{k \in I(\underline{y}_j)} s_{\mathbf{x}'\mathbf{y}'}(k) - 1$ to $y'_j = \underline{y}_j$. Considering that $x_{k^*} < \underline{y}_j < y_{l^*}$, there is no element in the initial joint sample $v \in \{x_1, \ldots, x_n, y_1, \ldots, y_m\}$ satisfying $x_{k^*} < v < \underline{y}_j$. Thus, the new rank assignment can be defined as follows:

$$s_{\mathbf{x}'\mathbf{y}'}(t) = \begin{cases} s_{\mathbf{xy}}(k^*) + 1 & \text{if } t = i, \\ s_{\mathbf{xy}}(k^*) & \text{if } t = j + n, \\ s_{\mathbf{xy}}(t) - 1, & \text{if } t \in \{1, \ldots, n\} \setminus \{i\} \text{ and } x_i < x_t \leq x^*_k, \\ s_{\mathbf{xy}}(t) + 1, & \text{if } t \in \{1+n, \ldots, m+n\} \setminus \{j\} \text{ and } y_{l^*} < y_{t-n} < y_j. \end{cases}$$

It is easy to check that the value of the Wilcoxon rank-sum statistic associated with the new rank assignment is equal to $r+1$, under any of the above possible situations. □

2.2. Bounds for the p-value and decision procedure

According to [1–3,5] we should proceed as follows, in order to make a decision from our imprecise dataset, **XY**. First, we need to calculate the set of possible values for the p-value, according to the imprecise information provided by **XY**:

$$P_{\mathbf{XY}} = \{p_{\mathbf{xy}} : \mathbf{xy} \in \mathbf{XY}\}.$$

Based on the information provided by the above set, and for a specific significance level, α, when the resulting bounds are on one side of α, the decision of the test hypothesis is clear. But when the bounds straddle the threshold, the test is inconclusive, since the data imprecision prevents us from making a clear determination. Thus, we will make one of the following decisions:

$$D(\mathbf{XY}) = \begin{cases} \text{reject}, & \text{if } P_{\mathbf{XY}} \subseteq [0, \alpha] \\ \text{accept}, & \text{if } P_{\mathbf{XY}} \subseteq (\alpha, 1] \\ \text{no-decision}, & \text{if } P_{\mathbf{XY}} \cap [0, \alpha] \neq \emptyset \text{ and } P_{\mathbf{XY}} \cap (\alpha, 1] \neq \emptyset. \end{cases} \tag{2}$$

For the particular test considered in this paper, the p-value associated with an arbitrary joint sample $\mathbf{xy} \in \mathbf{XY}$, and a specific rank assignment, $s_{\mathbf{xy}} \in \Sigma_{\mathbf{xy}}$, is determined as follows:

$$p(s_{\mathbf{xy}}) = 2\left[1 - \phi\left(\frac{|w(s_{\mathbf{xy}}) - \mu_0|}{\sigma_0}\right)\right],$$

where μ_0 and σ_0 respectively denote the expectation and standard deviation of the statistic under the null hypothesis. According to Theorem 2.1, the minimum and maximum of the set of values $\{t = \frac{w-\mu_0}{\sigma_0} : w \in W_{\mathbf{XY}}\}$ are respectively $\underline{t} = \frac{\min W_{\bar{\mathbf{x}}\underline{\mathbf{y}}} - \mu_0}{\sigma_0}$ and $\bar{t} = \frac{\max W_{\bar{\mathbf{x}}\underline{\mathbf{y}}} - \mu_0}{\sigma_0}$. Furthermore, we can easily check that, for an arbitrary interval $[a, b]$, the set of absolute values of the elements of it, $\{|x| : a \leq x \leq b\}$, coincides with the interval: $[\max\{0, -b, a\}, \max\{b, -a\}]$. Now, while taking into account that the distribution function ϕ is strictly increasing, we observe that the p-value increases as the absolute value of $t = \frac{w-\mu_0}{\sigma_0}$, with $w \in W_{\mathbf{XY}}$, decreases. Therefore, the set of possible p-values, $P_{\mathbf{XY}}$ is included in the interval $[\underline{p}, \bar{p}]$, whose extremes are determined as follows:

$$\underline{p} = 2\big[1 - \phi\big(\max\{\bar{t}, -\underline{t}\}\big)\big]$$
$$\bar{p} = 2\big[1 - \phi\big(\max\{0, -\bar{t}, \underline{t}\}\big)\big].$$

We can therefore easily check that the bound $\underline{p}$ is always attained for at least one of the extreme samples $\underline{\mathbf{x}}\bar{\mathbf{y}}$ and $\bar{\mathbf{x}}\underline{\mathbf{y}}$. Furthermore, when $\text{sign}(\underline{t}) = \text{sign}(\bar{t})$ (both positive or both negative) the upper bound $\bar{p}$ is also attained in at least one of those extreme samples. The only case where $\bar{p}$ is not reached is when $\underline{t} < 0 < \bar{t}$ and, furthermore, $\mu_0 \notin W_{\mathbf{XY}}$. In such situations, $\bar{p} = 2[1 - \phi(0)] = 1$, but $\max P_{\mathbf{XY}} = 2[1 - \phi(t^*)]$, where $|t^*| = \min\{|t| = |\frac{w-\mu_0}{\sigma_0}| : w \in W_{\mathbf{XY}}\} > 0$. Notwithstanding, we can derive from Theorem 2.2 that the distance between two consecutive values of $t = \frac{w(s_{\mathbf{xy}})-\mu_0}{\sigma_0}$, when we rank the whole imprecise joint sample, is equal to $\frac{1}{\sigma_0} = \sqrt{\frac{12}{nm(n+m+1)}}$. Under the conditions considered in this paper ($n \geq 10$ and

$m \geq 10$), this value is lower than 0.076. Therefore, we can state that the set $W_{\mathbf{XY}}$ contains at least one element w satisfying the inequality $|\frac{w-\mu_0}{\sigma_0}| \leq 0.076$, and so we can state that $\max P_{\mathbf{XY}} \geq 2[1-\phi(0.076)] \approx 0.94$. This value is higher than any significance level used in practice. According to this, replacing the actual set of possible p-values, $P_{\mathbf{XY}}$, by the interval $[\underline{p}, \overline{p}]$ calculated above will not modify our final decision in any practical situations, and will very much simplify the necessary calculations.

An alternative version of this test is provided in the literature: Instead of determining the value of $W_{\mathbf{xy}}$, the difference $W_{\mathbf{xy}} - W_{\mathbf{yx}}$ can be calculated. When taking into account the deterministic relation between $W_{\mathbf{xy}}$ and $W_{\mathbf{yx}}$ ($W_{\mathbf{xy}} + W_{\mathbf{yx}} = \sum_{i=1}^{n+m} i = \frac{(n+m)(n+m+1)}{2}$, we easily check that such a difference can be alternatively written as $2W_{\mathbf{xy}} - 2\mu_0$. Under the null hypothesis, this new statistic follows normal distribution with mean $\mu_0' = 0$, and standard deviation $\sigma_0' = 2\sigma_0 = \frac{nm(n+m+1)}{6}$. According to this new version of the test, the null hypothesis is rejected whenever the absolute value of the above difference is sufficiently large. Our algorithm returns the minimum and maximum possible values for $W_{\mathbf{xy}}$, i.e. $\min W_{\underline{\mathbf{x}}\overline{\mathbf{y}}}$ and $\max W_{\overline{\mathbf{x}}\underline{\mathbf{y}}}$. Following a similar procedure, but exchanging the roles of X and Y, we can apply it to calculate the minimum and the maximum values for $W_{\mathbf{yx}}$, that can be denoted by $\min W_{\underline{\mathbf{y}}\overline{\mathbf{x}}}$ and $\max W_{\overline{\mathbf{y}}\underline{\mathbf{x}}}$, respectively. Note however that there is also a deterministic relation between the above statistics. In fact, we can easily check that:

$$\min W_{\underline{\mathbf{x}}\overline{\mathbf{y}}} + \max W_{\overline{\mathbf{y}}\underline{\mathbf{x}}} = \max W_{\overline{\mathbf{x}}\underline{\mathbf{y}}} + \min W_{\underline{\mathbf{y}}\overline{\mathbf{x}}} = \frac{(n+m)(n+m+1)}{2}.$$

When taking this relation into account, we can easily check that the difference between the above intervals (according to set-valued arithmetic) provides the interval of possible values for the statistic $W_{\mathbf{xy}} - W_{\mathbf{yx}} = 2W_{\mathbf{xy}} - 2\mu_0$.

The above considerations concern the two-sided Wilcoxon rank sum test. In case of the one-sided variant, calculation of the maximum and minimum of the set of possible p-values, $P_{\mathbf{XY}}$ is simpler. In fact, if we consider the alternative hypothesis which states that $\mathbf{x}$ is shifted to the left of $\mathbf{y}$, the p-value would be written as follows:

$$p(s_{\mathbf{xy}}) = \phi\left(\frac{w(s_{\mathbf{xy}}) - \mu_0}{\sigma_0}\right).$$

Thus, the maximum and minimum possible values of the set $P_{\mathbf{XY}}$, given the information provided by imprecise sample, would be:

$$\underline{p} = \phi(\underline{t}) \quad \text{and} \quad \overline{p} = \phi(\overline{t}).$$

Similarly, the maximum and the minimum possible values of the set $P_{\mathbf{XY}}$, when we consider the other one-side test (H_1: $\mathbf{x}$ is shifted to the right of $\mathbf{y}$), can be calculated as follows:

$$\underline{p} = 1 - \phi(\overline{t}) \quad \text{and} \quad \overline{p} = 1 - \phi(\underline{t}).$$

We should finally make a decision according to Eq. (2).

3. The test in practice

3.1. Algorithm

Computing the generalized Wilcoxon rank-sum test is really easy and can be done at low computational cost compared to the original test that needs to find and process the links. This generalized test just involves two sorts, in case of precise data, and four sorts, in case of imprecise data. Since precise data is just a particular case of imprecise data, here we present the two algorithms needed to process imprecise data. Algorithm 1 computes $\min W_{\underline{\mathbf{x}}\overline{\mathbf{y}}}$ while Algorithm 2 computes $\max W_{\overline{\mathbf{x}}\underline{\mathbf{y}}}$.

Algorithm 1: Computation of $\min W_{\underline{\mathbf{x}}\overline{\mathbf{y}}}$.

Input: $\underline{\mathbf{x}}$, $\overline{\mathbf{y}}$, n, m,
Output: $\underline{W} = \min W_{\underline{\mathbf{x}}\overline{\mathbf{y}}}$
$\underline{W} = 0$;
set $\mathbf{z}$ = $[\underline{\mathbf{x}}, \overline{\mathbf{y}}]$ (concatenation) ;
$p = 1$;
sort $\mathbf{z}$; sort $\underline{\mathbf{x}}$;
for $i = 1, \ldots, (n+m)$ **do**
  if $\mathbf{z}(i) = \underline{\mathbf{x}}(p)$ **then**
    $p = p + 1$;
    $\underline{W} = \underline{W} + i$;

Algorithm 2: Computation of $\max W_{\bar{\mathbf{x}}\underline{\mathbf{y}}}$.

Input: $\bar{\mathbf{x}}$, $\underline{\mathbf{y}}$, n, m,
Output: $\overline{W} = \max W_{\bar{\mathbf{x}}\underline{\mathbf{y}}}$
$\overline{W} = 0$;
set $\mathbf{z} = [\bar{\mathbf{x}}, \underline{\mathbf{y}}]$ (concatenation) ;
$p = 1$;
sort $\mathbf{z}$; sort $\underline{\mathbf{y}}$;
for $i = 1, \ldots, (n+m)$ **do**
  if $\mathbf{z}(i) = \underline{\mathbf{y}}(p)$ **then**
    $p = p + 1$;
  else
    $\overline{W} = \overline{W} + i$;

3.2. A simple example

To clarify the computation of the generalized Wilcoxon rank-sum test for interval data, we propose, in this Section, a very simple example consisting of comparing two interval-valued samples $\mathbf{X} = ([1,3],[2,3],[3,5])$ and $\mathbf{Y} = ([2,3],5)$. Within this example $n = 3$, $m = 2$, $\mu_0 = \frac{3(3+2+1)}{2} = 9$ and $\sigma_0^2 = \frac{3\times 2(3+2+1)}{12} = 3$.

Computation of $\underline{W}$ according to Algorithm 1 requires defining the sorted vectors $\mathbf{z} = (1,2,3,3,5)$ and $\underline{\mathbf{x}} = (1,2,3)$. It outputs $\underline{W} = 1 + 2 + 3 = 6$.

Computation of $\overline{W}$ according to Algorithm 2 requires defining the sorted vectors $\mathbf{z} = (2,3,3,5,5)$ and $\underline{\mathbf{y}} = (2,5)$. It outputs $\overline{W} = 2 + 3 + 5 = 10$.

Then $\underline{t} = \frac{\underline{W}-\mu_0}{\sigma_0} = \frac{6-9}{\sqrt{3}} = -\sqrt{3}$ and $\bar{t} = \frac{\overline{W}-\mu_0}{\sigma_0} = \frac{10-9}{\sqrt{3}} = \frac{1}{\sqrt{3}}$. Thus, $\underline{p} = 2(1-\phi(\max\{\frac{1}{\sqrt{3}}, \sqrt{3}\})) = 2(1-\phi(\sqrt{3})) \approx 2(1-0.9584) \approx 0.083$ and $\bar{p} = 2(1-\phi(\max\{0, -\frac{1}{\sqrt{3}}, -\sqrt{3}\})) = 2(1-\phi(0)) = 2(1-0.5) = 1$.

In that case, we obtain an imprecise p-value which is highly imprecise. However, even if this interval is very wide, it completely falls in the acceptance region if we consider the usual 0.05-significance level test.

3.3. Power and type I error

This simulated experiment aims at comparing the proposed generalization of the Wilcoxon test with the precise original test in terms of power. The proposed experiment is based on considering a sample of n elements $(x_1, \ldots, x_n)$ taken from a uniform distribution on the interval $[-2,2]$ and a sample of m elements $(y_1, \ldots, y_m)$ from another uniform distribution $U(\mu-2, \mu+2)$, for a fixed μ in the interval $[-0.5, 0.5]$. The respective sample sizes satisfy the constraints $n + m = 1000$ and $400 \leq n \leq 600$.

We repeated this experiment 1000 times, and we determined the rejection rates in the precise Wilcoxon test, in order to obtain an estimation of the power of the Wilcoxon rank-sum classical test, $\text{pow}(\mu)$, for each particular value of $\mu \neq 0$, as well as the type I error for $\mu = 0$. Fig. 2 shows these rates in blue for different values of α (in particular, we have considered the significance levels $\alpha = 0.01$, $\alpha = 0.05$, and $\alpha = 0.1$ – plotted from top to bottom). We determined those rates for every $\mu = -0.5 + \frac{i}{100}$, $i = 0, 1, \ldots, 100$ (values of μ are displayed on the x-axis on every graph). We have added imprecision to those samples according to the following procedure. For each particular level of imprecision $\Delta > 0$, and each sample value x_i, we constructed an interval $[\underline{x_i}, \bar{x_i}]$ of length Δ centered on x_i. The same level of imprecision, Δ was considered on the values of the second sample. The process has been repeated for the values $\Delta = 0.2$, $\Delta = 0.1$ and $\Delta = 0.05$ (from left to right in the figures). Fig. 1 shows the rejection, "acceptance" and "no-decision" rates, according to the generalized imprecise test. We observed that the rejection rate was an increasing function with respect to the absolute value of μ, while the acceptance rate decreased with respect to it. The "no-decision" rate was low when the absolute value of μ was very high (and thus the rejection rate was high) or very close to 0 (where the acceptance rate was high). We compared the different plots from the right to the left and observed that the no-decision rate increased with Δ. We compared the different plots from top to bottom and observed that the rejection rate increased while the acceptance rate decreased with α. Fig. 2 compares the rejection rates and the rates of "no acceptance" (rejection or no-decision) of the generalized test to the respective rejection rates of the classical test. For each particular value of μ, Δ and α, each pair of rates provides a pair of estimations of the upper and the lower bounds of the "power" of the test, based on the respective imprecise datasets. According to Fig. 2, the distance between the rejection rate (in red) and the non-acceptance rate (in black) increased with respect to Δ, i.e. the degree of imprecision. We also observed that both rates increased with respect to α (the greater the significance level, the lower the acceptance rate).

Let us take into account that in our simulated dataset, the "true" precise values coincide with the center of the intervals in the imprecise dataset. Similar plots can be obtained under the less restrictive situation where the center of the intervals does not necessarily coincide with those "true" precise values, but rather with their expectations. If, on the contrary, our imprecise data are biased to the left or to the right, the upper and lower rejection rates would change (depending on the

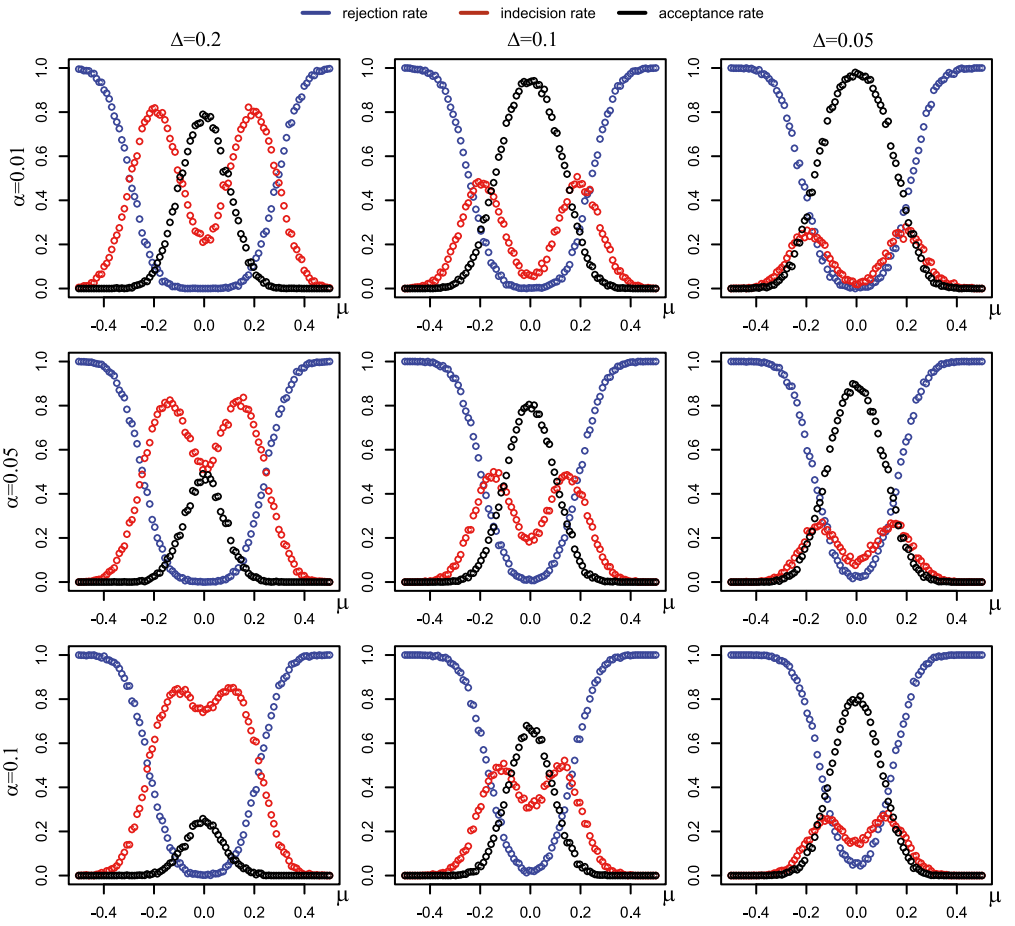

Fig. 1. "Rejection", "no-decision" and "acceptance" rates wrt generalized Wilcoxon imprecise test. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

different values of μ). Notwithstanding, something that would not change is that the difference between both rates would still increase with respect to the imprecision degree, Δ.

3.4. Experiments

A major issue concerning statistical tests is their robustness with respect to the data model. The Wilcoxon rank-sum test was designed to compare samples taken from pairs of continuous distributions, but it is frequently applied to pairs of samples taken from discrete distributions. In particular, it is often used to deal with quantized information, which involves comparing pairs of integer vectors.

For example, pixel values of a 512×512 gray-level image are quantized on 8 bits, i.e. on 256 values. On average, 1024 pixels would thus have the same value. Using a Wilcoxon rank-sum test and ignoring the quantization effect could lead to arbitrary results, and thus wrong decisions: the value of the statistic calculated from quantized data could differ markedly from a calculation based on original (non-quantized) data.

In some applications, e.g. image based medical diagnosis, inferential statistics are often used to compare pairs of regions in the same image or in two different images, and determine whether they are similar or not. Early diagnosis is vital in the treatment of some diseases, which means that we need to detect differences at an early stage. In such cases, accepting the equality hypothesis when it is false must be avoided as much as possible. On the contrary rejecting this hypothesis when it is true may sometimes lead to an unnecessary harmful treatment choice. In fact, physicians would usually try to avoid a wrong decision, and prefer to acquire additional data when the actual data are not fully reliable. Thus, knowing that no-decision can be taken based on the current set of data is a valuable piece of information.

Finally, when the degree of sensor imprecision is known, it can also be valuable to be able to use this information in the test.

We propose two illustrative experiments that highlight the ability of the generalized Wilcoxon rank-sum test to avoid wrong decisions. Those experiments are based on images acquired by a gamma camera (nuclear medicine images). They aim

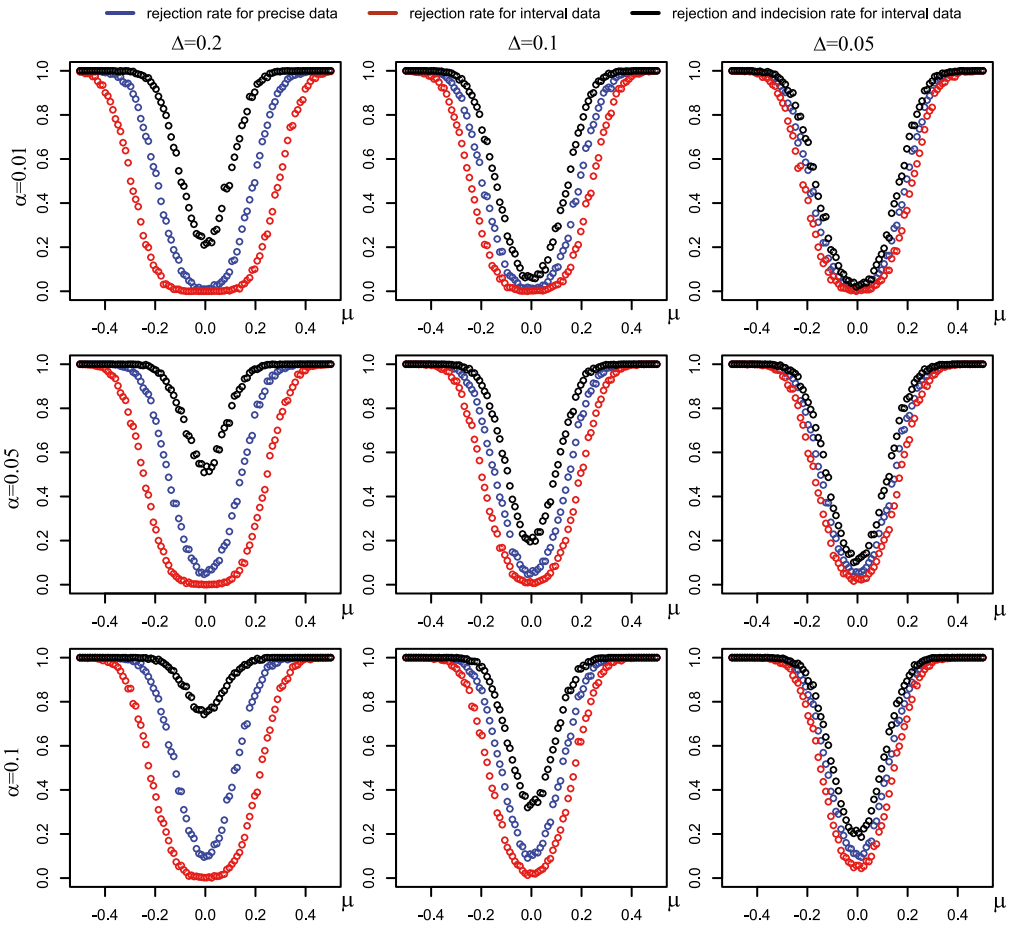

Fig. 2. Comparison between generalized imprecise Wilcoxon test and the classical one. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

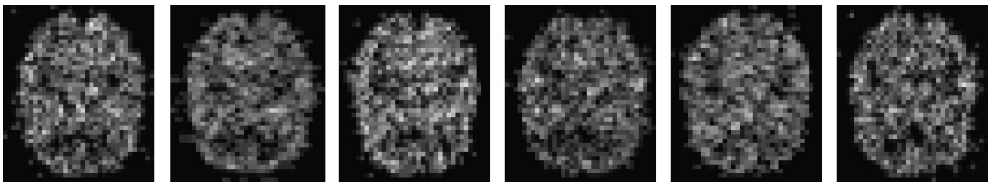

Fig. 3. 6 acquisitions of the Hoffman 2-D brain phantom.

at mimicking real medical situations where the nuclear physician has to compare the distribution of values in two regions of interest in order to decide whether or not a patient has a specific disease.

3.4.1. Material

A Hoffman 2-D brain phantom (Data Spectrum Corporation) was filled with a 99*m* technetium solution (148 MBq/L) and placed on one of the detectors of a dual-head gamma camera using a low-energy high-resolution parallel-hole collimator (INFINIA, General Electric Healthcare). A dynamic study was carried out to obtain 1000 planar acquisitions (acquisition time: 1 second; average count per image 1.5 kcounts, 128×128 images to satisfy the Shannon condition), representing 1000 measures of a random 2-D image (see Fig. 3). The Hoffman phantom is designed to simulate, by a partial volume effect, the emissivity ratio between brain gray matter and white matter. A gamma camera counts the number of photons that have been emitted in a particular direction. Thus, pixel values in a nuclear image are counts and therefore can be assumed to be contaminated by Poisson distributed noise.

As the acquisition time was very short, the images were very noisy, i.e. the signal to noise ratio was very low. More precisely, the average pixel value in the brain corresponded to a 69% coefficient of variation in the Poisson noise. Moreover, due to this short time acquisition, the number of different possible values to be assigned to a pixel was low

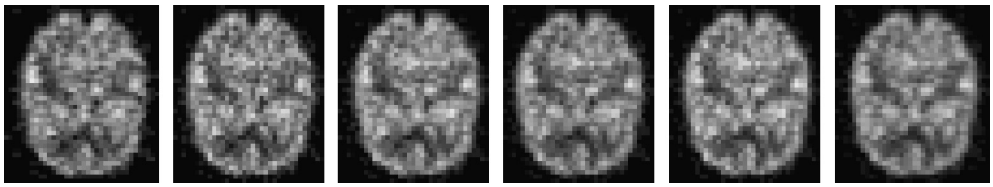

Fig. 4. 6 images obtained by summing up 10 raw acquisitions of the Hoffman 2-D brain phantom.

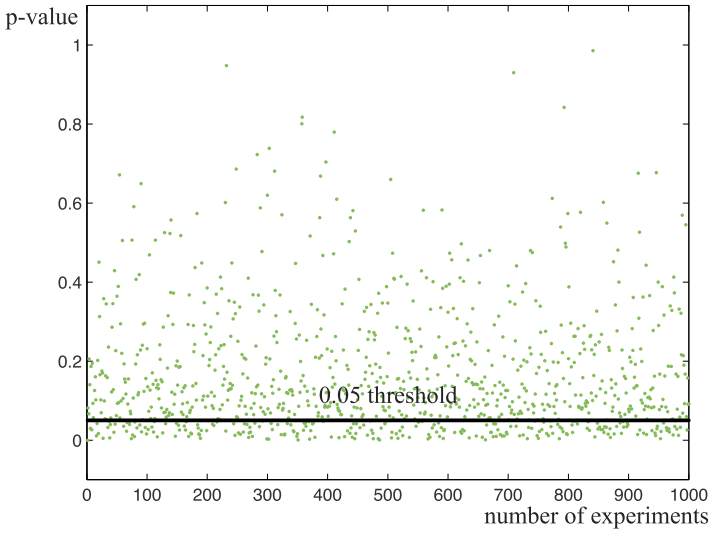

Fig. 5. Classical Wilcoxon rank-sum test obtained by comparing 1000 pairs of raw acquisitions of the Hoffman 2-D brain phantom.

and thus, within those images, the quantization impact was high. For example, in raw images, the pixel values were $\{0, 256, 512, 768, 1024, 1280, 1536, 1792, 2048\}$. This quantization was induced by the gamma-camera technology, but could also be considered as a quantized values of a Poisson-ruled process associated with each pixel.

To obtain less noisy and less quantized images, we summed the raw images (see e.g. Fig. 4). The higher the number of summed images, the higher the average pixel value, and thus the higher the signal to noise ratio. When summing the 1000 raw images, we obtained the high dynamic resolution and high signal to noise ratio image depicted in Fig. 10a.

3.4.2. The generalized test accounts for quantization

Due to the short acquisition time, the raw images were quantized on nine levels: $0, 256, 512, 768, 1024, 1280, 1536, 1792, 2048$. The impact of quantization was therefore substantial. To highlight the effect of quantization on the Wilcoxon rank-sum test, we considered 1000 acquisitions of the same image. We formed 1000 pairs, each of them containing each of the 1000 images and another image taken at random from the remaining 999 images. We compared the pixel value distributions in every pair of those raw images. More precisely, we compared the vector of the non-null pixels of the upper half of the first image to the vector of the non-null pixels of the lower half of the second image. We compared both distributions, without considering the pixel locations. Of course, other kinds of image comparisons could be done, but this is beyond the scope of this illustrative example. The number of non-null pixels differed between images and therefore the sample sizes n and m did not generally coincide but were always greater than 20. In this experiment, the data were considered as precise. The distributions in the upper and lower parts were almost identical and in fact were identical when considering the high dynamic image depicted in Fig. 10a. The test results should thus not lead us to reject the null hypothesis that states that both distributions are identical. In fact, the proportion of times that the p-value is below a pre-defined threshold α is expected to be around α. Figs. 5 and 6 plot the p-values respectively obtained using the classical and generalized Wilcoxon rank-sum test. The usual threshold value of $\alpha = 0.05$ was superimposed on each figure (in black). The p-values displayed in Fig. 5 should have corresponded to a sample of size 1000 taken from a uniform distribution on the unit interval. Thus, only 5% of those values (around 50 times out of 1000) would be expected to be lower than $\alpha = 0.05$. However, they were below this threshold 253 times out of 1000 ($> 25\%$), leading to a wrong decision (type I error). This paradoxical result (the wrong decision ratio should be under 0.05), was due to the fact that the calculation of the Wilcoxon rank-sum test critical values was based on quantized data rather than continuous data.

Conversely, the p-value provided by the generalized Wilcoxon rank-sum test was very imprecise: the upper p-value, $(\overline{p})$ equaled 1 while the lower p-value, $(\underline{p})$, equaled 0 for any instance of the test. It shows that, according to this test, no-decision could be made with these data, since there were too many repeated values in the quantized sample. In a

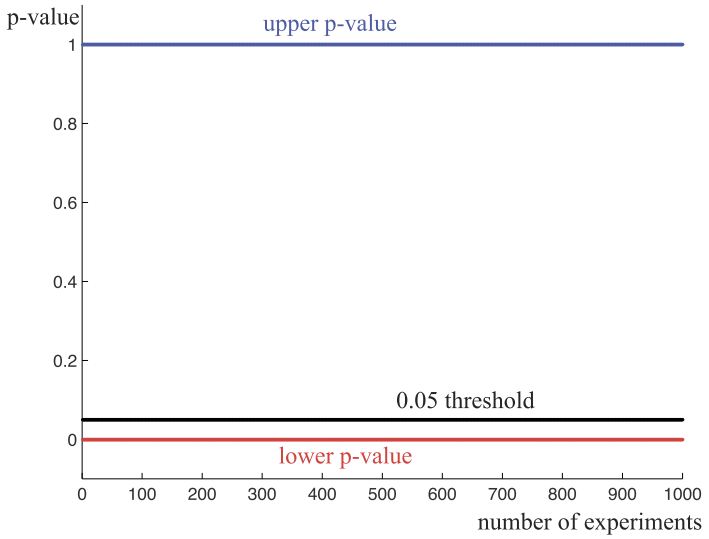

Fig. 6. Generalized Wilcoxon rank-sum test obtained by comparing 1000 pairs of raw acquisitions of the Hoffman 2-D brain phantom.

practical case, this situation would indicate to the physician that a new data acquisition would be required with a higher acquisition time.

3.4.3. The generalized test accounts for imprecision

In this second experiment, we considered images with a higher signal to noise ratio obtained by summing the pixel values, pixel by pixel, after taking a fixed number k acquisitions of the same image. We thus obtained 500 images when summing pairs of raw images, 333 images when summing triplets of images, etc. When the signal to noise ratio increased, the data were quantized on a higher number of values and thus the noise and quantization effects decreased. For example, summing up two raw images led to an image quantized on 14 levels: 0, 256, 512, 768, 1024, 1280, 1536, 1792, 2048, 2304, 2560, 2816, 3072 (values 3584 and 4096 could appear in theory, but not in practice).

For each signal to noise ratio (i.e. for each value of k), we achieved 100 comparisons between the distribution of (summed) pixel values in the lower part of one image and the distribution of the (summed) pixel values in the upper part of another image. We then counted the number of times where the test led to a wrong decision (i.e. times where it led us to conclude that both distributions were different). For the classical Wilcoxon rank-sum test, it corresponded to the number of times where the p-value went under 0.05. For the generalized Wilcoxon rank-sum test, it corresponded to the number of times when the upper p-value went under 0.05. We also counted, in this last case, the number of times where the generalized test was inconclusive, because the bounds of the p-value straddled the threshold (i.e. $\underline{p} < 0.05 < \overline{p}$). We then focused on the ratios (i.e. number of counts over the total number of considered pairs of images). Fig. 7 plots those ratios versus the number of summed images (k).

Note that the ratio of wrong decisions when considering the generalized test was lower than the ratio of wrong decisions when considering the classical test. The proportion of times that the generalized test was inconclusive was also very high. As expected, every ratio converged at 0.05 when k (and thus the signal to noise ratio) increased. We also noted a very paradoxical phenomenon: when k increased from 2 to 20, the number of wrong decisions increased in both cases. This was due to the combined effect of random noise, quantization and the fact that non-null pixel values increased with summation, leading to a decision that the two distributions differed even though they did not.

Although the wrong decision ratio was low when using the generalized test, it was still non-null, and therefore did not fully reflect the ability of this test to avoid wrong decisions.

In a second stage, we considered each pixel value to be Poisson distributed, with mean λ (with λ depending on the pixel location). We thus artificially made the data imprecise by replacing each value x by a confidence interval for λ centered on this single observation x. We opted for the confidence levels $1-\alpha = 0.95$ and $1-\alpha = 0.68$, often used in statistics, because of their relation with the 1σ and 2σ intervals for Gaussian distributions. Each confidence interval $I_\alpha(x)$ was computed according to the classical formula $I_\alpha(x) = [F^{-1}(\frac{\alpha}{2}, x, 1), F^{-1}(1-\frac{\alpha}{2}, x+1, 1)]$, where $F^{-1}(\cdot, n, 1)$ is the quantile function of the gamma distribution with a shape parameter n and scale parameter 1. The result is plotted in Fig. 8 for $1-\alpha = 0.95$ and Fig. 9 for $1-\alpha = 0.68$.

When considering imprecise data, the wrong decision ratio was lower, but the no-decision ratio increased, compared to the result when considering precise data. Moreover, as noted in Section 3.3, the higher the data imprecision, the higher the no-decision ratio (the no-decision ratio increases with the confidence level of the intervals). This experiment highlighted the ability of this new test to account for known imprecision in the data and to transform it into imprecision in the decision. This test could allow physicians to calibrate the acquisition time needed to avoid making a wrong decision.

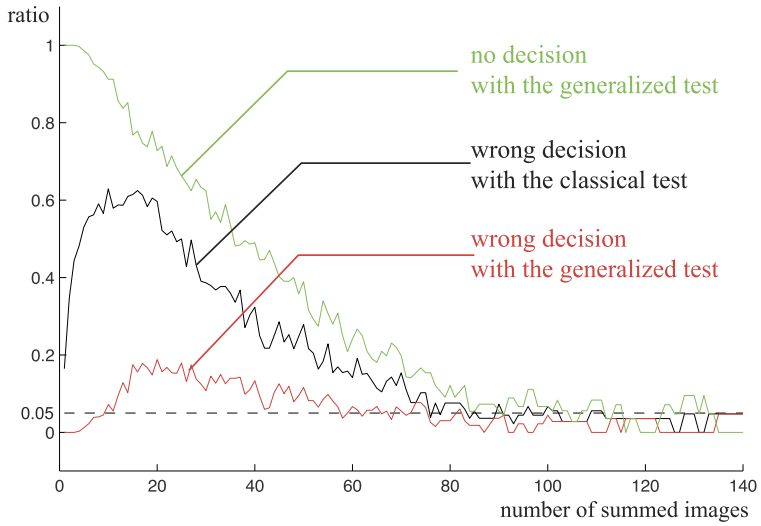

Fig. 7. Test with precise data: wrong decision ratio with the classical test (black) and with the generalized test (red), superimposed with the proportion of times it was inconclusive (green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

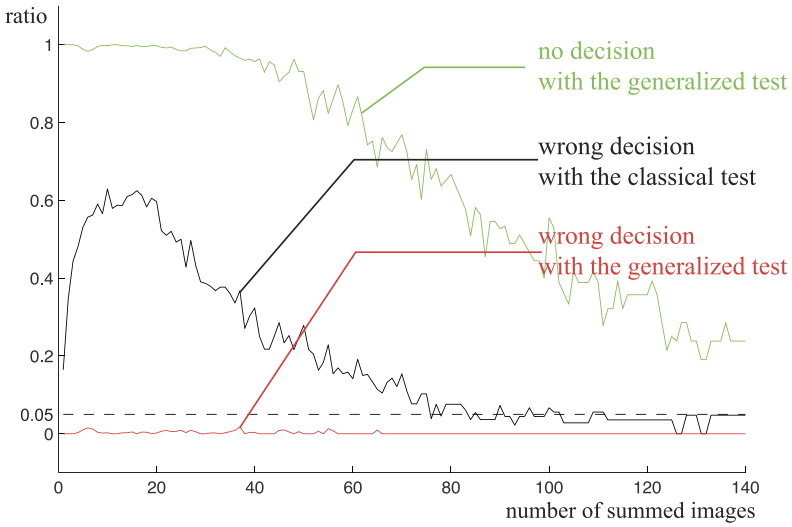

Fig. 8. Test with imprecise data (95% confidence intervals): wrong decision ratio with the classical test (black) and with the generalized test (red), superimposed with the proportion of times it was inconclusive (green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Remark. The generalized test with imprecise data does not always lead to imprecise decisions.

To illustrate this, we asked the nuclear medicine physician to manually select, on the high dynamic image obtained by summing the 1000 raw images (Fig. 10a), the pixels that trustfully belonged to brain white matter (Fig. 10b) and gray matter (Fig. 10c). Instead of comparing the upper and lower non-null regions in pairs of images, we compared the gray matter pixels of an image with the white matter pixels of the other image. The distributions of pixel values differed in those cases, so therefore a type II error would occur if the null hypothesis is not rejected.

We achieved this comparison in 200 pairs of images and counted the number of times the white matter distribution in an image seemed to match the gray matter distribution in the other image (wrong decisions), both for the classical test and generalized test. For the generalized test, we also counted the number of times that it was inconclusive for different signal to noise ratios. For the lowest signal to noise ratio ($k = 1$), the classical test provided 1.5% of wrong decisions, while the generalized test was conclusive 33.5% times, and among those times, it always made the right decision (rejecting the null hypothesis). As soon as $k > 1$, the classical test provided no wrong decisions and the generalized test was always conclusive, and it made the right decision in every instance. In other words, both tests concluded that both distributions differed, in every instance. The results are the same when considering imprecise or precise valued data.

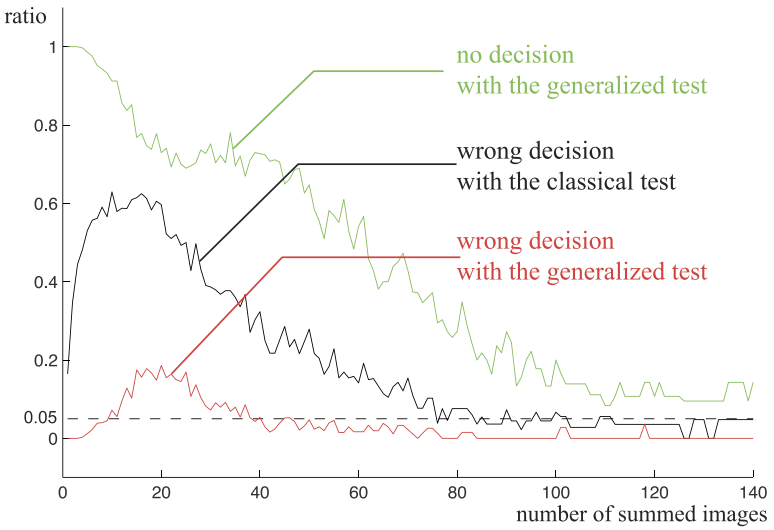

Fig. 9. Test with imprecise data (68% confidence intervals): wrong decision ratio with the classical test (black) and with the generalized test (red), superimposed with the proportion of times it was inconclusive (green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

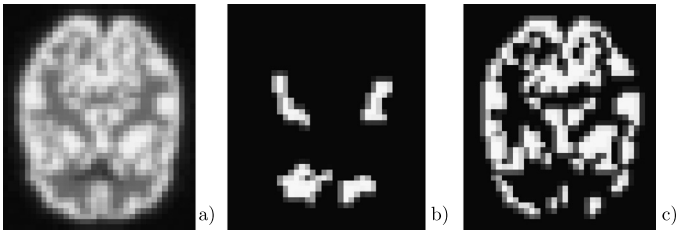

Fig. 10. Reference image obtained by summing the 1000 raw images (a), pixels belonging to the white matter (b) and to the gray matter (c).

4. Conclusion

The Wilcoxon rank-sum test was designed to check whether two independent real valued samples correspond to the same distribution or not. This test must be adjusted to deal with quantized or imprecise data. In this paper, we proposed such an adaption by generalizing the classical test. One of the main features of this new test is that it provides a pair of bounds, $\underline{p}$ and $\overline{p}$, for the p-value (instead of a precise number) leading to a bipolar decision: the answer to such a test can be *yes*, *no* or *unknown*. We proved that the interval $[\underline{p}, \overline{p}]$ we computed contains all p-values that should have been obtained by using the conventional test with precise valued data belonging to the set of interval-valued data. This new test also deals with the quantization in a new way. Quantization is perceived as a non-linear modification of the supposed existing true real samples that disturb the test. When the disturbance is too marked, the test is inconclusive. If the test leads to a decision, the test level can be guaranteed.

As a future work, we consider extending this test to fuzzy interval-valued data which would lead to a fuzzy interval of p-values. A very straightforward and exact solution may be obtained by using Algorithms 1 and 2 for each level-cut of the fuzzy interval-valued data, and building the level cut of the p-value fuzzy interval. This solution is tractable only if the membership values are quantized (i.e. can take only a finite number of values). Otherwise, it would lead to a too computationally expensive algorithm. Finding a low computational cost algorithm for extending the generalized Wilcoxon rank-sum test remains an open problem.

Acknowledgements

The authors would like to thank Dr. Denis Mariano Goulart (Nuclear Medicine Department of the Montpellier Lapeyronie Hospital) for providing the data that was used in the experimental section, and for his helpful remarks and comments. This work was partially supported by the Spanish Project TIN2011-24302 and the INSERM Physicancer 2012 – PC201211 – STIPPI project.

References

[1] I. Couso, L. Sánchez, Defuzzification of fuzzy p-values, in: D. Dubois, et al. (Eds.), Soft Methods for Handling Variability and Imprecision, in: Advances in Soft Computing, vol. 48, 2008, pp. 126–132.
[2] I. Couso, L. Sánchez, Mark–recapture techniques in statistical tests for imprecise data, Int. J. Approx. Reason. 52 (2011) 240–260.
[3] T. Denœux, M.-H. Masson, P.-A. Hébert, Nonparametric rank-based statistics and significance tests for fuzzy data, Fuzzy Sets Syst. 153 (2005) 1–28.
[4] S. Ferson, V. Kreinovich, J. Hajagos, L. Oberkampf, W. Ginzburg, Experimental uncertainty estimation and statistics for data having interval uncertainty, Unlimited release Sandia National Laboratories, 2007, SAND2007–0939.
[5] P. Filzmoser, R. Viertl, Testing hypotheses with fuzzy data: the fuzzy p-value, Metrika 59 (2004) 21–29.
[6] O. Hryniewicz, Goodman–Kruskal γ measure of dependence for fuzzy ordered categorical data, Comput. Stat. Data Anal. 51 (2006) 323–334.
[7] O. Hryniewicz, K. Opara, Efficient calculation of Kendall's ? for interval data, in: R. Kruse (Ed.), Synergies of Soft Computing and Statistics for Intelligent Data Analysis, 2013, pp. 203–210.
[8] L. Jaulin, M. Kieffer, E. Walter, D. Meizel, Guaranteed robust nonlinear estimation with application to robot localization, IEEE Trans. Syst. Man Cybern., Part C, Appl. Rev. 32 (2003) 374–382.
[9] L. Jaulin, E. Walter, Guaranteed nonlinear parameter estimation from bounded-error data via interval analysis, Math. Comput. Simul. 35 (1993) 123–127.
[10] J. Otero, L. Sanchez, I. Couso, A. Palacios, Bootstrap analysis of multiple repetitions of experiments using an interval-valued multiple comparison procedure, J. Comput. Syst. Sci. 80 (2014) 88–100.
[11] J. Otero, L. Sánchez, I. Couso, A. Palacios, Bootstrap analysis of multiple repetitions of experiments using an interval-valued multiple comparison procedure, J. Comput. Syst. Sci. 80 (2014) 88–100.
[12] A. Palacios, L. Sánchez, I. Couso, Diagnosis of dyslexia with low quality data with genetic fuzzy systems, Int. J. Approx. Reason. 51 (2010) 993–1009.
[13] I. Perfilieva, V. Novák, A. Dvořák, Fuzzy transform in the analysis of data, Int. J. Approx. Reason. 48 (18) (2008) 36–46.
[14] L. Sanchez, I. Couso, Advocating the use of imprecisely observed data in genetic fuzzy systems, IEEE Trans. Fuzzy Syst. 15 (2007) 551–562.
[15] O. Strauss, A. Rico, Towards interval-based non-additive deconvolution in signal processing, Soft Comput. 16 (5) (2012) 809–820.
[16] M. Černý, J. Antoch, M. Hladík, On the possibilistic approach to linear regression models involving uncertain, indeterminate or interval data, Inf. Sci. 244 (2013) 26–47.
[17] F. Wilcoxon, Individual comparisons by ranking methods, Biom. Bull. 1 (1945) 80–83.
[18] S. Xu, Q. Luo, G. Xu, L. Zhang, Asymmetrical interval regression using extended ϵ-svm with robust algorithm, Fuzzy Sets Syst. 160 (2009) 988–1002.