sciendo

Int. J. Appl. Math. Comput. Sci., 2024, Vol. 34, No. 2, 277–289
DOI: 10.61822/amcs-2024-0020

amcs

BOOTSTRAPPED TESTS FOR EPISTEMIC FUZZY DATA

PRZEMYSŁAW GRZEGORZEWSKI [a], MACIEJ ROMANIUK [b,*]

[a]Faculty of Mathematics and Information Science
Warsaw University of Technology
Koszykowa 75, 00-662 Warsaw, Poland
e-mail: przemyslaw.grzegorzewski@pw.edu.pl

[b]Systems Research Institute
Polish Academy of Sciences
Newelska 6, 01-447 Warsaw, Poland
e-mail: mroman@ibspan.waw.pl

Epistemic bootstrap is a resampling algorithm that generates bootstrap real-valued samples based on some epistemic fuzzy data input. We apply this method as a universal basis for various statistical tests which can be then directly used for fuzzy random variables. Two classical goodness-of-fit tests are considered as an example to examine the suggested methodology for both synthetic and real data. The proposed approach is also compared with two other goodness-of-fit tests dedicated directly to fuzzy data.

Keywords: bootstrap, fuzzy data, nonparametric statistics, simulation, statistical computing.

1. Introduction

The bootstrap, as introduced by Efron (1979), is a recognized and widely used statistical inference tool. However, due to some disadvantages, various extensions of it have been proposed for real-valued data, like the smoothed bootstrap, the wild bootstrap, etc. (De Angelis and Young, 1992; Chernick *et al*., 2011). The bootstrap was also applied to fuzzy data (e.g., Gil *et al*., 2006; González-Rodríguez *et al*., 2006; Montenegro *et al*., 2004), but some pitfalls similar to those for real data, like the frequent repetition of several values for small samples, were still visible. Therefore, a few new resampling methods for fuzzy data have been proposed recently (Grzegorzewski *et al.* 2019; 2020a; 2020b; Romaniuk and Hryniewicz, 2021), oriented, however, towards *ontic fuzzy data* (Couso and Dubois, 2014), i.e., the data that appear as essentially fuzzy-valued. But, in practice, *epistemic data* (Couso and Dubois, 2014) are also widespread. They refer to some exact data which exist objectively, but they are imprecisely observed so their true real values remain unknown. The concept of *epistemic data* is more natural, intuitive, and appealing, e.g., in engineering, where imprecise measurements are common.

First attempts for applying the bootstrap to epistemic fuzzy data were discussed by Grzegorzewski and Romaniuk (2021; 2022a; 2022b). In this paper, we proceed to fill the gap between the "bootstrap world" and the "epistemic fuzzy data universe". Our new contribution is threefold. Firstly, we generalize the epistemic bootstrap so that it can be applied to a wide class of nonparametric statistical tests (not only for the Kolmogorov–Smirnov test as in our previous work (Grzegorzewski and Romaniuk, 2022b)). Secondly, an entirely new resampling algorithm leading to the final conclusion (i.e., to accept/reject the null hypothesis) is provided. This algorithm is related to the so-called boot-perm test. Finally, through a comprehensive and detailed numerical analysis, we compare the introduced approach applied for the Kolmogorov–Smirnov test and the Cramér-von Mises test with the respective "classical" (i.e., real-valued) counterparts. Besides small and moderate samples of synthetic data, real-life case studies are considered. Our epistemic bootstrap algorithms are also compared with two other tests for fuzzy data known in the literature (Grzegorzewski, 2020; Grzegorzewski

*Corresponding author

and Gadomska, 2021). Some of the introduced methods are available in the R package *FuzzySimRes* (Romaniuk *et al*., 2023). Our approach is a very general one, contrary to others, aimed at selected statistical tests or only some families of probability distributions (e.g., Hesamian *et al*., 2023).

The paper is organized as follows. Section 2 delivers preliminaries concerning fuzzy data modeling. Two general epistemic bootstrap algorithms are proposed in Section 3. Next, in Section 4, four variants of epistemic bootstrapped goodness-of-fit tests for fuzzy data are discussed. The suggested methods are examined through a simulation study in Section 5. Concluding remarks are given in Section 6.

2. Preliminaries

In practice, unfortunately, instead of exact results of experiments, we are dealing with their imprecise measurements or vague perceptions. This is especially the case when the so-called human factor plays a decisive role in obtaining the results. An appropriate mathematical model of such imprecise data is then necessary to allow drawing reliable conclusions despite the lack of access to precise data. In the case of statistical inference, it is also necessary to adapt traditional procedures in such a way that it is possible to use them to analyze imprecise data or to construct new statistical tools. A convenient environment for modeling imprecision, as well as for further work of data analysts, is the theory of fuzzy sets. In the case of experimenters whose real results are real numbers (vectors), a special subfamily of fuzzy sets, i.e., fuzzy numbers, turns out to be of interest.

Definition 1. A mapping $\widetilde{x} : \mathbb{R} \to [0,1]$ is a *fuzzy number* if its α-cuts $(\widetilde{x})_\alpha$, defined by

$$(\widetilde{x})_\alpha = \begin{cases} \{x \in \mathbb{R} : \widetilde{x} \geq \alpha\} & \text{if } \alpha \in (0,1], \\ cl\{x \in \mathbb{R} : \widetilde{x} > 0\} & \text{if } \alpha = 0, \end{cases}$$

are nonempty compact intervals for all $\alpha \in [0,1]$. The operator cl stands here for the closure.

Thus, in the remainder of this paper, we will deal with situations where, instead of real-valued outcomes of experiments $\mathbb{x} = (x_1, \ldots, x_n)$, where $x_i \in \mathbb{R}$, we have a respective sequence of fuzzy numbers $\widetilde{\mathbb{x}} = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

A fuzzy number as a real function, even with the constraints resulting from the definition, can take various shapes. The most commonly used fuzzy numbers are the so-called *trapezoidal fuzzy numbers* given by

$$\widetilde{x}(x) = \begin{cases} \frac{x-a}{b-a} & \text{if } a < x \leq b, \\ 1 & \text{if } b \leq x \leq c, \\ \frac{d-x}{d-c} & \text{if } c \leq x < d, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $a, b, c, d \in \mathbb{R}$, and $a \leq b \leq c \leq d$. Since each such fuzzy number is completely described by four real values, we can denote a trapezoidal fuzzy number by $\widetilde{x} = \mathrm{Tra}(a, b, c, d)$. If $b = c$, then $\widetilde{x}$ is a *triangular fuzzy number*. Trapezoidal (triangular) fuzzy numbers are so popular to use because of their simplicity in processing and calculations, together with their natural interpretation. This is also the reason why fuzzy numbers with more complicated shapes are often approximated by trapezoidal fuzzy numbers (Grzegorzewski, 2008; Ban *et al*., 2015).

Further on, the family of all fuzzy numbers will be denoted by $\mathbb{F}(\mathbb{R})$, while the family of trapezoidal fuzzy numbers by $\mathbb{F}^T(\mathbb{R})$.

If our imprecise data are the result of some random experiment, their generation mechanism can be related to fuzzy-valued random variables (Kwakernaak, 1978; Kruse, 1982).

Definition 2. Given a probability space $(\Omega, \mathcal{F}, P)$, a mapping $\widetilde{X} : \Omega \to \mathbb{F}(\mathbb{R})$ is said to be a *fuzzy random variable* (f.r.v.) if, for each $\alpha \in [0,1]$, $(\inf \widetilde{X}_\alpha) : \Omega \to \mathbb{R}$ and $(\sup \widetilde{X}_\alpha) : \Omega \to \mathbb{R}$ are real-valued random variables on $(\Omega, \mathcal{F}, P)$.

In this case, a fuzzy random variable $\widetilde{X}$ might be considered a *fuzzy perception* of a "standard" (but unknown) random variable X, called the *original* of $\widetilde{X}$. Similarly, a whole *fuzzy random sample* $\widetilde{\mathbb{X}} = (\widetilde{X}_1, \ldots, \widetilde{X}_n)$ is a fuzzy perception of a usual real-valued random sample.

3. Epistemic bootstrap

Any statistical inference depends on the underlying population, a sampling procedure, and the statistics, i.e., a function $T = T(\mathbb{X})$, relevant to the problem considered. All these factors provide the so-called sampling distribution. Sometimes it can be easily identified, e.g., if our sample consists of n independent observations from the same normal distribution and we want to verify the null hypothesis about its mean, the required test statistic is t-distributed with $n-1$ degrees of freedom. In statistical practice, we often encounter situations when the population distribution is unknown, so the sampling distribution cannot be designed straightforwardly. But if the sample size is large enough, we can often make inference based on the asymptotic distribution. Otherwise, we apply appropriate distribution-free procedures. However, if we are dealing with imprecise data, we also encounter difficulties of another nature.

Suppose that we have a fuzzy sample $\widetilde{\mathbb{x}} = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$ that is an imprecise perception of the real-valued sample $\mathbb{x} = (x_1, \ldots, x_n)$. Based on the epistemic view, we may say that $\widetilde{x}_i$ contains the actual real-valued realization of the i-th observation, but

we do not have information about its precise location. Clearly, this entails additional difficulties in identifying the sampling distribution. Fortunately, the membership function of $\widetilde{x}_i$ provides the necessary knowledge about the possibility that every single point is the true realization of X_i. Anyway, achieving the goal requires a very large amount of data, while our sample is not always very large. And this is where Effron's idea of the bootstrap to create new random samples based on what is given to us comes in very useful. Obviously, to take into account our lack of knowledge due to the data imprecision, another type of bootstrap that works differently than the one proposed by Effron is necessary. Its idea, called the *epistemic bootstrap*, was introduced and developed by Grzegorzewski and Romaniuk (2021; 2022a; 2022b).

The idea of their two-step resampling procedure is quite simple (see Algorithm 1). In the first step, given the initial fuzzy sample $\widetilde{\mathbb{x}} = \widetilde{x}_1, \ldots, \widetilde{x}_n$, we generate randomly a value α from the uniform distribution on the unit interval (denoted as $\mathrm{U}[0,1]$) for each $i = 1, \ldots, n$. This value points out the respective level of the α-cut for $\widetilde{x}_i$. In the second step, a real value x_i^* from the selected α-cut of $\widetilde{x}_i$ is independently drawn from the uniform distribution on this α-cut (denoted as $\mathrm{U}\big((\widetilde{x}_i)_{\alpha_i}\big)$). In this way, we obtain a new real-valued bootstrap sample $x_1^*, \ldots, x_n^*$. The advantage of the bootstrap is that we are not limited to one sample, but we can multiply any number of them, so we usually generate $B \geq 1$ bootstrap samples denoted by $\mathbb{x}_j^* = (x_{1j}^*, \ldots, x_{nj}^*)$, where $j = 1, \ldots, B$. This procedure is known as the standard (or simple) epistemic bootstrap. Its new version, called the *antithetic approach*, was proposed by Grzegorzewski and Romaniuk (2022a) to improve statistical efficiency. In this case, the second step was slightly altered—besides drawing a single real value x_i' from the selected α-cut of $\widetilde{x}_i$, the additional real value x_i'' is generated from its counterpart, i.e., $(1-\alpha)$-cut, using the respective uniform distribution $\mathrm{U}\big((\widetilde{x}_i)_{1-\alpha_i}\big)$. Then these two values are averaged to give the output x_i^* (see Algorithm 2).

Algorithm 1. Epistemic fuzzy bootstrap: the standard approach.

Require: Initial fuzzy sample $\widetilde{x}_1, \ldots, \widetilde{x}_n \in \mathbb{F}(\mathbb{R})$.
Ensure: B bootstrap samples.
for $j = 1$ to B **do**
  for $i = 1$ to n **do**
    $\alpha_{ij} \leftarrow \mathrm{U}[0,1]$.
    $x_{ij}^* \leftarrow \mathrm{U}\big((\widetilde{x}_i)_{\alpha_{ij}}\big)$.
  end for
end for
return $\mathbb{x}_j^* = (x_{1j}^*, \ldots, x_{nj}^*)$, where $j = 1, \ldots, B$.

Algorithm 2. Epistemic fuzzy bootstrap: the antithetic approach.

Require: Initial fuzzy sample $\widetilde{x}_1, \ldots, \widetilde{x}_n \in \mathbb{F}(\mathbb{R})$.
Ensure: B bootstrap samples.
for $j = 1$ to B **do**
  for $i = 1$ to n **do**
    $\alpha_{ij} \leftarrow \mathrm{U}[0,1]$.
    $x_{ij}' \leftarrow \mathrm{U}\big((\widetilde{x}_i)_{\alpha_{ij}}\big)$.
    $x_{ij}'' \leftarrow \mathrm{U}\big((\widetilde{x}_i)_{1-\alpha_{ij}}\big)$.
    $x_{ij}^* = \frac{1}{2}\big(x_{ij}' + x_{ij}''\big)$.
  end for
end for
return $\mathbb{x}_j^* = (x_{1j}^*, \ldots, x_{nj}^*)$, where $j = 1, \ldots, B$.

4. Epistemic bootstrapped test

Epistemic bootstrap sample generation, while essential and interesting in itself, is not so much a goal as an initial step to perform statistical inference of fuzzy data. After introducing both algorithms, we will move on to the second step, showing how resampled data can be used for goodness-of-fit testing in the two-sample problem. It will also be an illustration of how classical statistical tools can be transferred to the fuzzy domain.

Let $\widetilde{\mathbb{X}} = (\widetilde{X}_1, \ldots, \widetilde{X}_n)$ and $\widetilde{\mathbb{Y}} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_m)$ denote fuzzy random samples taken independently from two populations with unknown cumulative distribution functions (c.d.f.) F and G, respectively. Our goal is to check whether F and G differ significantly, or we can assume that both samples come from the same distribution. More formally, we are interested in verifying the following null hypothesis:

$$H_0 : F(t) = G(t) \quad \text{for all } t \in \mathbb{R} \tag{2}$$

against the alternative $H_1 : F(t) \neq G(t)$ (for some $t \in \mathbb{R}$).

If the available samples consist of real-valued data, one can apply many tests for verifying H_0, like the Kolmogorov–Smirnov test (Smirnov, 1933) (abbreviated further as the KS test) or the Cramer–von Mises test (Anderson, 1962) (CvM test for short), among the best known regarding this issue. We will use both of these tests to show how the epistemic bootstrap can be used to analyze fuzzy data.

Let $\widetilde{\mathbb{x}} = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$ and $\widetilde{\mathbb{y}} = (\widetilde{y}_1, \ldots, \widetilde{y}_m)$ stand for the actual realization of our fuzzy samples. Following Algorithms 1 or 2 we generate B epistemic bootstrap samples $\mathbb{x}_1^*, \ldots, \mathbb{x}_B^*$ and $\mathbb{y}_1^*, \ldots, \mathbb{y}_B^*$, each consisting of n or m real-valued observations, respectively.

Let us imagine again, for a moment, that we are dealing with the classical KS or CvM test based on real-valued data. Since both tests have right-sided critical regions, we compute the corresponding p-value as

$$p = \mathbb{P}_{H_0}\left(T \geq t\right), \tag{3}$$

where $t = T(\mathbb{x}, \mathbb{y})$ is the actual value of the test statistic.

Returning to the fuzzy data, thanks to the epistemic bootstrap, we now have B pairs of samples, so we can derive multiple test statistic values. However, to decide on the null hypothesis, we can: (a) aggregate the values of these statistics and compute the p-value following (3), or (b) find a p-value for each statistic separately and then aggregate the p-values. All in all, we can follow four different ways described below.

Multi-statistic approach. Given B pairs of the epistemic bootstrap samples $(\mathbb{x}_j^*, \mathbb{y}_j^*)$, $j = 1, \ldots, B$, we can compute B corresponding values of the test statistic considered, i.e., $t_1^* = T(\mathbb{x}_1^*, \mathbb{y}_1^*), \ldots, t_B^* = T(\mathbb{x}_B^*, \mathbb{y}_B^*)$, and next, following (3), the respective p-values

$$p_j^* = \mathbb{P}_{H_0}\left(T \geq t_j^*\right), \quad j = 1, \ldots, B. \tag{4}$$

Then these p-values should be combined into a single p-value $p^* = A(p_1^*, \ldots, p_B^*)$, where A is some aggregation function. Many examples of such functions have been proposed in the literature, but most of them require independence between tests, which we cannot assume in the case under consideration. Thus, without assuming any particular dependence structure among the p-values, the Simes method (Simes, 1986) is usually recommended (and we apply it in our analysis), so that

$$A(p_1^*, \ldots, p_B^*) = \min_{k=1,\ldots,B} \frac{B}{k}\, p_{(k)}, \tag{5}$$

where $p_{(k)}$ is the k-th smallest p-value among $p_1^*, \ldots, p_B^*$. Another natural way is to combine these p-values using their average (at least when the tests have similar power), i.e., $\overline{p}^* = \sum_{j=1}^{B} p_j^*$. Unfortunately, $\overline{p}^*$ is not necessarily a p-value. For more information on combining p-values, we refer the reader to Vovk and Wang (2020).

Resampling approach. If the null hypothesis H_0 holds, then it should not matter how we pair the samples when determining the value of the statistic T. Therefore, given the epistemic bootstrap samples $\mathbb{x}_1^*, \ldots, \mathbb{x}_B^*$ and $\mathbb{y}_1^*, \ldots, \mathbb{y}_B^*$, we can use additional resampling to select randomly K pairs $(\mathbb{x}_{k_1}^*, \mathbb{y}_{k_2}^*)$, where k_1 and k_2 are randomly and independently picked up with probability $1/B$ from the set $1, \ldots, B$. Then we compute the p-values as follows:

$$p_k^* = \mathbb{P}_{H_0}\left(T \geq T(\mathbb{x}_{k_1}^*, \mathbb{y}_{k_2}^*)\right), \quad k = 1, \ldots, K. \tag{6}$$

Finally, these p-values have to be aggregated (e.g., with the Simes method). A natural question appears of why to use not all possible but only $K \leq B$ pairs. Obviously, if B is small, there is no problem with considering all possible pairs of samples, but if B is large, we reduce the numbers of the pairs considered due to the numerical efficiency.

Averaging approach. The methods discussed earlier require p-value aggregation, which is usually controversial although necessary in some applications. In our case, this can be avoided by using the typical bootstrap approach, i.e., by averaging the statistics obtained in subsequent steps. Namely, we can determine a desired single p-value by computing

$$p^* = \mathbb{P}_{H_0}(T \geq t^{**}), \tag{7}$$

where

$$t^{**} = \frac{1}{B}\sum_{j=1}^{B} t_j^* = \frac{1}{B}\sum_{j=1}^{B} T(\mathbb{x}_j^*, \mathbb{y}_j^*). \tag{8}$$

Otherwise, we can combine the last formula with the resampling approach applied for obtaining $t_j^* = T(\mathbb{x}_{k_1}^*, \mathbb{y}_{k_2}^*)$.

Bootperm approach. As the name suggests, it consists of two basic steps combining two methods: bootstrap and permutation tests. Suppose we have two initial fuzzy samples $\widetilde{\mathbb{x}}$ and $\widetilde{\mathbb{y}}$. In the first step (bootstrap), we generate the respective epistemic bootstrap samples $\mathbb{x}^*$ and $\mathbb{y}^*$ and calculate our test statistic value $t^* = T(\mathbb{x}^*, \mathbb{y}^*)$. Then, in the step typical for permutation tests, the two samples are pooled into one, i.e., we consider $\mathbb{w}^* = \mathbb{x}^* \uplus \mathbb{y}^*$, where $\uplus$ stands for the vector concatenation, so that $w_i^* = x_i^*$ if $1 \leq i \leq n$ and $w_i^* = y_i^*$ if $n+1 \leq i \leq n+m$. A key premise for a further action is that, if H_0 holds, then all available observations may be viewed as if they were randomly assigned to both samples, but they come from the same population. Therefore, we create a permutation $\mathbb{w}^{**}$ of $\mathbb{w}^*$ and divide it into two subsamples by assigning first n elements of $\mathbb{w}^{**}$ to $\mathbb{x}^{**}$ and the remaining m elements into $\mathbb{y}^{**}$. Next we determine the test statistic value $t^{**} = T(\mathbb{x}^{**}, \mathbb{y}^{**})$. After considering such K permutations, we obtain $t_1^{**}, \ldots, t_K^{**}$ which can be used to approximate the p-value (see Algorithm 3):

$$p_j^{**} = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}(t_k^{**} > t^*). \tag{9}$$

Finally, by repeating the whole procedure B times, we receive $p_1^{**}, \ldots, p_B^{**}$, which are then aggregated to give p^*.

5. Simulation study

To compare the proposed methods, we conduct a broad simulation study. We restricted our experiments to trapezoidal fuzzy numbers, which dominate in the practical use. Moreover, as was shown by Lubiano *et al.* (2017), the shape of the membership function scarcely affects statistical conclusions. To construct a fuzzy number $\widetilde{x} = \mathrm{Tra}(a, b, c, d)$, as given by (1), we need four real numbers. Therefore, we generate fuzzy samples as

Algorithm 3. Boot-perm goodness-of-fit test.

Require: Initial fuzzy samples $\widetilde{\mathbb{x}}$ and $\widetilde{\mathbb{y}}$.
Ensure: The approximated p-value p_j^{**}.
1: Generate epistemic bootstrap samples $\mathbb{x}^* = (x_1^*, \ldots, x_n^*)$ and $\mathbb{y}^* = (y_1^*, \ldots, y_m^*)$ based on $\widetilde{\mathbb{x}}$ and $\widetilde{\mathbb{y}}$.
2: Calculate $t^* = T(\mathbb{x}^*, \mathbb{y}^*)$.
3: Pool the data $\mathbb{w}^* = \mathbb{x}^* \uplus \mathbb{y}^*$.
4: **for** $k = 1$ to K **do**
5: Draw a permutation $\mathbb{w}^{**}$ of $\mathbb{w}^*$.
6: Divide $\mathbb{w}^{**}$ into two samples $\mathbb{x}^{**}$ and $\mathbb{y}^{**}$.
7: Calculate $t_k^{**} = T(\mathbb{x}^{**}, \mathbb{y}^{**})$.
8: **end for**
9: Compute p_j^{**} using (9).
10: **return** p-value p_j^{**}.

realizations of trapezoidal fuzzy random numbers defined as follows:

$$a = X - S^l - C^l, \qquad b = X - C^l, \qquad c = X + C^r, \qquad d = X + C^r + S^r, \tag{10}$$

where X is a random variable corresponding to the "true" population distribution, while C^l, C^r, S^l, S^r denote random variables used for "bluring" X and producing its fuzzy perception $\widetilde{x}$ (Grzegorzewski *et al.*, 2019; 2020b; Grzegorzewski and Romaniuk, 2022a; Romaniuk and Hryniewicz, 2021; Romaniuk and Grzegorzewski, 2023). Two random variables C^l and C^r are applied to create the core of $\widetilde{x}$ to avoid naive defuzzification identifying the original value with the center of the core. Additionally, S^l and S^r are used for modeling the support of the generated fuzzy observation. All these random variables are generated independently from the distributions given in Table 1. The notation in the table is self-explanatory, e.g., if $\widetilde{x}_i \in \mathbb{F}_{(\mathrm{N,U,U})}$, then X is simulated from the standard normal distribution with zero mean and unit standard deviation, C^l, C^r are generated from the uniform distribution on the interval $(0, 0.6)$, and S^l, S^r from the uniform distribution on $(0, 0.8)$. Similarly, $\mathrm{Exp}(\lambda)$ denotes the exponential distribution with the parameter λ, $\Gamma(\alpha, \beta)$ – the gamma distribution with the shape parameter α and scale parameter β, and $\beta(a, b)$ stands for the beta distribution with the parameters a, b, etc. The first two models are used to analyze a difference in location and dispersion (see Sections 5.1 and 5.2), while the others—to compare a difference in distributions' shape (Section 5.3).

In our experiments, we considered samples of relatively small ($n = m = 10$) or moderate ($n = m = 100$) sizes. Firstly, we performed the classical KS and CvM tests (in the second case, the *CvM2SL1Test* package (Xiao, 2012) was used) for the real-valued samples $\mathbb{X} = (X_1, \ldots, X_n)$ and $\mathbb{Y} = (Y_1, \ldots, Y_m)$. Next, following (10), fuzzy samples $\widetilde{\mathbb{x}}$ and $\widetilde{\mathbb{y}}$ were generated. Then B epistemic bootstrap samples $\mathbb{x}^*$ and $\mathbb{y}^*$ are created from these fuzzy samples using the standard (see Algorithm 1) or the antithetic (see Algorithm 2) method. Finally, the epistemic bootstrapped versions of the KS and CvM tests are conducted (abbreviated as EKS and ECvM, respectively). Actually, four versions of each bootstrapped test, corresponding to the methods proposed in Section 4, were considered. To shorten the notation, we use the following: *ms* for the multi-statistic approach, *res* for the statistics resampling, *avs* for the averaging approach, *btp* for the boot-perm approach, and *std* for the standard and *ant* for the antithetic methods. The Simes method (from the *metapod* package (Lun, 2021)) or plain averaging is applied to aggregate the p-values.

In our study, we know the actual distribution of real-valued samples. Although such a situation usually does not occur in practice, here it allows us to compare our epistemic versions of tests with their classical counterparts used as reference points. To reduce randomness, each experiment was repeated 10000 times and the outcomes were averaged. To evaluate the results the following measures related to the confusion matrix are also calculated: accuracy (ACC), false positive ratio (FP), and false negative ratio (FN). ACC is the frequency of situations when both the epistemic test (*ET*) and its classical counterpart (*CT*) lead to the same decision concerning the null hypothesis at the significance level $\alpha_{sl} = 0.05$, i.e.,

$$\mathrm{ACC} = \frac{1}{l}\sum_{i=1}^{l}\Big(\mathbb{1}(\text{ET rejects } H_0)\cdot\mathbb{1}(\text{CT rejects } H_0) + \mathbb{1}(\text{ET accepts } H_0)\cdot\mathbb{1}(\text{CT accepts } H_0)\Big), \tag{11}$$

$$\mathrm{FP} = \frac{1}{l}\sum_{i=1}^{l}\mathbb{1}(\text{ET accepts } H_0)\cdot\mathbb{1}(\text{CT rejects } H_0), \tag{12}$$

$$\mathrm{FN} = \frac{1}{l}\sum_{i=1}^{l}\mathbb{1}(\text{ET rejects } H_0)\cdot\mathbb{1}(\text{CT accepts } H_0), \tag{13}$$

where $\mathbb{1}(\cdot)$ is the indicator function.

Only some numerical results are further provided. Other outcomes and figures are available upon request.

5.1. Detecting a difference in location. We examined firstly the test behavior in situations when the distributions differ only in location. It was enforced by the deterministic shift added to Y.

In the case of the EKS test and the initial samples from $\mathbb{F}_{(\mathrm{N,U,U})}$ and $\mathbb{F}_{(\mathrm{E,U,U})}$, the *ms* and *res* methods give bigger p-values, while the *avs* approach—lower p-values in comparison with the KS test (Fig. 1). For the small sample (i.e., $n = m = 10$), the *btp* approach leads to

Table 1. Scenarios for simulating fuzzy random variables.

Type	X	C^l, C^r	S^l, S^r
$\mathbb{F}_{(\mathrm{N,U,U})}$	$\mathrm{N}(0,1)$	$\mathrm{U}(0,0.6)$	$\mathrm{U}(0,0.8)$
$\mathbb{F}_{(\mathrm{E,U,U})}$	$\mathrm{Exp}(\frac{1}{2})$	$\mathrm{U}(0,0.6)$	$\mathrm{U}(0,1.2)$
$\mathbb{F}_{(\beta,\mathrm{U,U}),1}$	$\beta(2,5)$	$\mathrm{U}(0,0.6)$	$\mathrm{U}(0,0.8)$
$\mathbb{F}_{(\mathrm{U,U,U}),1}$	$\mathrm{U}(\frac{4-\sqrt{15}}{14}, \frac{4+\sqrt{15}}{14})$	$\mathrm{U}(0,0.6)$	$\mathrm{U}(0,0.8)$
$\mathbb{F}_{(\mathrm{N,U,U}),1}$	$\mathrm{N}\left(\frac{2}{7}, \sqrt{\frac{5}{196}}\right)$	$\mathrm{U}(0,0.6)$	$\mathrm{U}(0,0.8)$
$\mathbb{F}_{(\Gamma,\mathrm{U,U}),1}$	$\Gamma\left(\frac{16}{5}, \frac{5}{56}\right)$	$\mathrm{U}(0,0.6)$	$\mathrm{U}(0,0.8)$

very small p-values. This method behaves better for the bigger sample (i.e., $n = m = 100$) when the obtained p-values are still lower but comparable with their "crisp" counterparts. Surprisingly, averaging p-values (instead of obtaining the respective values by the Simes method) leads to significant improvements of the estimated final p-values, especially for the *ms* and *res* approaches (Fig. 4). Accuracy is usually very high (more than 80–90%; Fig. 2), apart from the *btp* approach, where some very low values (even about 30–40%) can be noticed. It seems that the decrease in ACC in the interval $[0.5, 1.5]$ is caused by the increase in both the FP and FN, but the FN contributes significantly more in this case.

Generally, the Simes method leads to larger p-values for all of the approaches, apart from the *btp* method. The differences for ACC are insignificant, also besides the *btp* method. The values of FP are usually rather low (about 5–6%), especially for the *avs* and the *btp* approaches (even about 1–2%). But the *btp* algorithm gives bigger values for the FN (even about 30%). It seems that the *ms-ant* and *res-ant* methods are the golden means regarding this criterion, and averaging p-values improves the results. The estimated power curves (Fig. 3) for the EKS tests are usually slightly higher than for the KS test, but averaging the p-values leads to the overlapping of the results, apart from the *btp* method with the questionably high values.

For the CvM/ECvM tests, the conclusions are very similar (Figs. 5 and 6). However, the *btp* approach leads to the p-values which are closer to the "crisp" outputs, even for the small sample, but ACC is still lower for the *btp* approach than for other methods. The power curve for the CvM test is slightly higher than its counterparts for the ECvM tests, apart from the *btp* method for which the values are clearly too big. The averaging of p-values gives very good approximations of the "crisp" p-values and lowers the distances between the power graphs. It can also lead to an almost negligible loss for ACC.

In general, it is advisable to use the *ant* method instead of its *std* counterpart, and the averaging of the obtained p-values may be profitable.

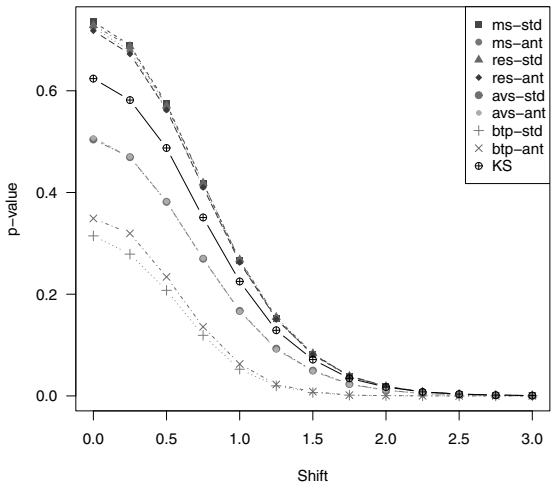

Fig. 1. Empirical p-values for the KS test, $X \sim \mathbb{F}_{(\mathrm{N,U,U})}$, and $K = 10, n = m = 10$ (shift added to Y).

5.2. Detecting a difference in dispersion. Next, we compared the tests' behavior in situations when both distributions differ only in dispersion. This was modeled by the gradually increasing variance of the second sample Y when the initial samples are from $\mathbb{F}_{(\mathrm{N,U,U})}$.

The results are similar to those obtained above, i.e., the *ms* and *res* approaches give bigger p-values, and *avs* —lower p-values when compared with the KS test. The *btp* approach leads to relatively low values for the small sample. However, for ECvM, *avs* is a promising approach (Fig. 7). The ACC levels (apart from the *btp* method) are usually very high (Fig. 8). The power curve for the KS test is slightly lower (or higher for the CvM test) when compared with its EKS (or ECvM) counterparts. The averaging of p-values can improve the final p-value, reduces distances between power curves, and has almost negligible effect for ACC (apart from the *btp* approach, which significantly lowers this value). Therefore, the *avs, ms-ant, res-ant* approaches should be preferred.

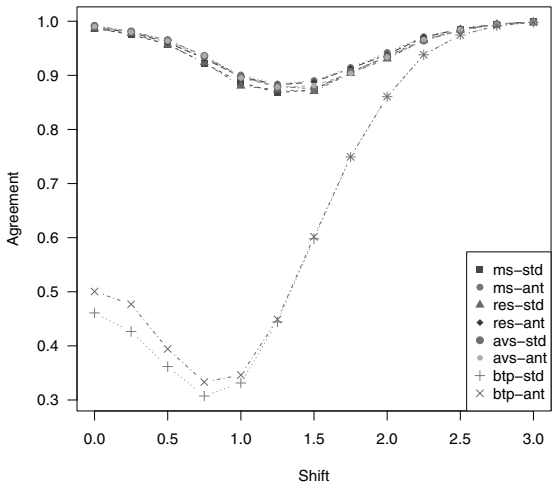

Fig. 2. Simulated ACC values for the KS test, $X \sim \mathbb{F}_{(\mathrm{N,U,U})}$, and $K = 10, n = m = 10$ (shift added to Y).

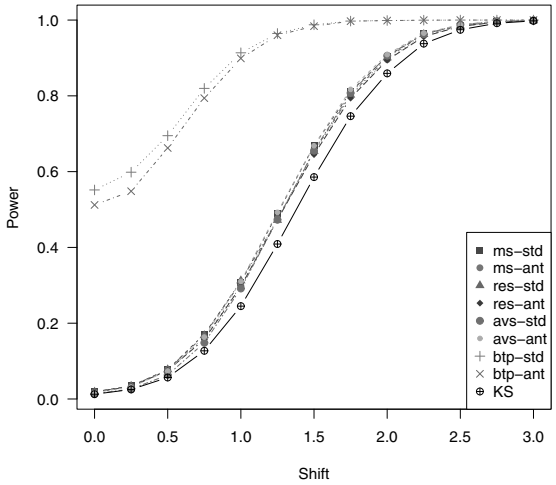

Fig. 3. Simulated power for the KS test, $X \sim \mathbb{F}_{(\mathrm{N,U,U})}$, and $K = 10, n = m = 10$ (shift added to Y).

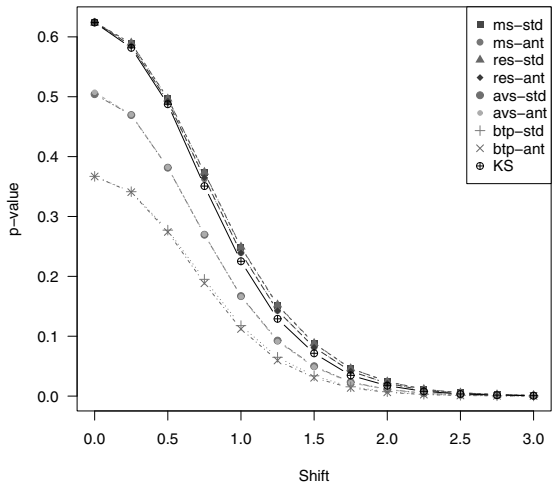

Fig. 4. Empirical p-values for the KS test, $X \sim \mathbb{F}_{(\mathrm{N,U,U})}$, and $K = 10, n = m = 10$ (shift added to Y), averaged p-values.

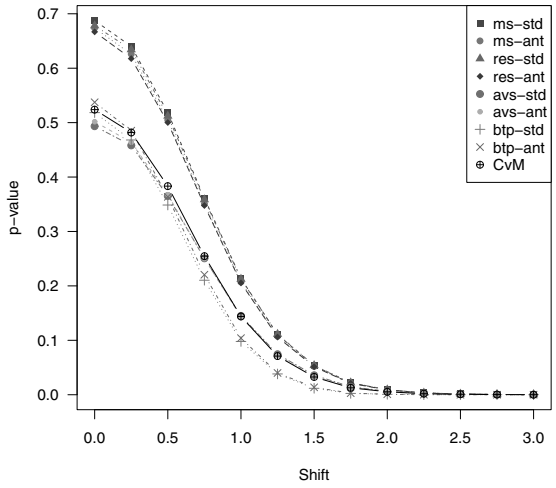

Fig. 5. Empirical p-values for the CvM test, $X \sim \mathbb{F}_{(\mathrm{N,U,U})}$, and $K = 10, n = m = 10$ (shift added to Y).

5.3. Detecting a difference in shape. In the next study, X and Y come from quite different distributions, but are selected to have identical expected values and variances. For instance, $X \sim \beta(2,5)$ and $Y \sim \mathrm{N}(2/7, \sqrt{5/196})$, where $\mathbb{E}X = \mathbb{E}Y = 2/7$ and $\mathrm{Var}X = \mathrm{Var}Y = 5/196$.

Results in Table 2 show that the *avs* and *ms* methods deliver p-values close to the results of the KS/CvM tests, while the *btp* approach gives relatively low p-values (to make it more readable, the p-values for the EKS/ECvM tests which are the most similar to the KS/CvM counterparts are given in bold). ACC is usually very high (about 80–90%), apart from the *btp* method (see Table 3). The *ant* approaches give better results than their *std* versions. Averaging p-values produces the aggregated p-values slightly closer to the KS/CvM counterparts and leads to a more significant improvement of ACC.

To better check the quality of our estimators, we estimated their standard errors. The obtained values were very low, e.g., for the p-values from Table 2, the estimated standard errors were equal from 0.0011 up to 0.0037, and for ACC given in Table 3, these values were in the interval 0.0021–0.0050. Then, it seems that our conclusions are not influenced by the variability caused by the randomness of our numerical experiments.

5.4. Real-life case study.

5.4.1. Lifetimes of street light equipment. In Hesamian and Taheri (2013), fuzzy data concerning the

Table 2. Comparison of the estimated p-values for the CvM test and different distributions.

	ms-std	ms-anti	res-std	res-ant	avs-std	avs-ant	btp-std	btp-ant	CvM
$\mathbb{F}_{(\beta,\mathrm{U},\mathrm{U}),1}$ vs $\mathbb{F}_{(\mathrm{U},\mathrm{U},\mathrm{U}),1}$									
$n=10$	0.6345	0.7000	0.6481	0.7066	0.4382	**0.4768**	0.0981	0.1535	0.5067
$n=100$	0.2524	0.6191	0.2793	0.6326	0.2638	0.4588	0.0392	**0.2815**	0.3091
$\mathbb{F}_{(\beta,\mathrm{U},\mathrm{U}),1}$ vs $\mathbb{F}_{(\mathrm{N},\mathrm{U},\mathrm{U}),1}$									
$n=10$	0.6319	0.6959	0.6484	0.7016	0.4373	**0.4749**	0.0971	0.1488	0.5134
$n=100$	0.2490	0.6152	0.2761	0.6245	0.2628	**0.4571**	0.0368	0.2820	0.3860
$\mathbb{F}_{(\beta,\mathrm{U},\mathrm{U}),1}$ vs $\mathbb{F}_{(\Gamma,\mathrm{U},\mathrm{U}),1}$									
$n=10$	0.6318	0.6982	0.6486	0.7048	0.4378	**0.4767**	0.0974	0.1470	0.5213
$n=100$	0.2419	0.6148	0.2689	0.6294	0.2604	**0.4610**	0.0366	0.2775	0.4337

Table 3. Comparison of ACC for the CvM test and different distributions.

	ms-std	ms-anti	res-std	res-ant	avs-std	avs-ant	btp-std	btp-ant
$\mathbb{F}_{(\beta,\mathrm{U},\mathrm{U}),1}$ vs $\mathbb{F}_{(\mathrm{U},\mathrm{U},\mathrm{U}),1}$								
$n=10$	0.9355	0.9403	0.9365	0.9419	0.9529	0.9531	0.2217	0.2977
$n=100$	0.7888	0.8896	0.7946	0.8879	0.8982	0.8990	0.1747	0.4581
$\mathbb{F}_{(\beta,\mathrm{U},\mathrm{U}),1}$ vs $\mathbb{F}_{(\mathrm{N},\mathrm{U},\mathrm{U}),1}$								
$n=10$	0.9309	0.9373	0.9325	0.9339	0.9454	0.9459	0.2261	0.2968
$n=100$	0.7890	0.8947	0.7952	0.8924	0.9054	0.9066	0.1604	0.4500
$\mathbb{F}_{(\beta,\mathrm{U},\mathrm{U}),1}$ vs $\mathbb{F}_{(\Gamma,\mathrm{U},\mathrm{U}),1}$								
$n=10$	0.9387	0.9439	0.9383	0.9422	0.9549	0.9551	0.2173	0.2864
$n=100$	0.7995	0.9205	0.8126	0.9255	0.9388	0.9401	0.1308	0.4280

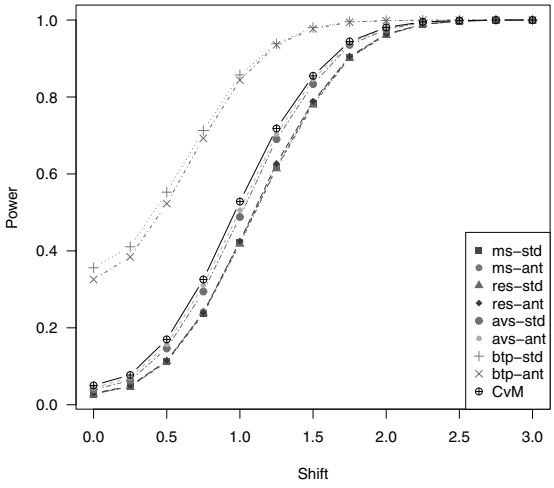

Fig. 6. Simulated power for the CvM test, $X \sim \mathbb{F}_{(\mathrm{N},\mathrm{U},\mathrm{U})}$, and $K = 10, n = m = 10$ (shift added to Y).

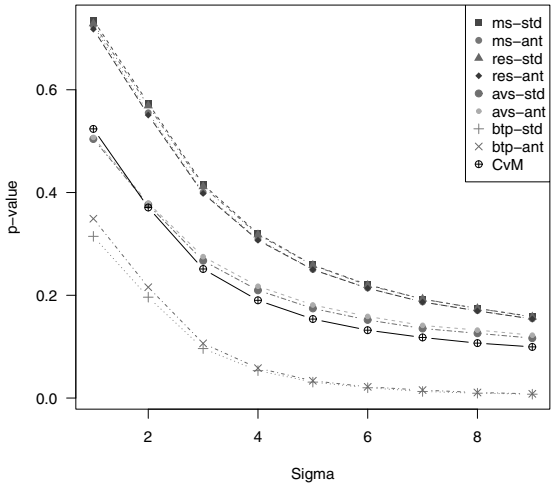

Fig. 7. Empirical p-values for the CvM test, $X \sim \mathbb{F}_{(\mathrm{N},\mathrm{U},\mathrm{U})}$, and $K = 10, n = m = 10$ (increasing σ).

lifetimes of street light equipment for two suppliers (denoted by A and B) are analyzed. We would like to check if these lifetimes can be regarded as "similar", i.e., originating from the same distributions of originals for these two samples. In the work of Gibbons and Chakraborti (2010), the respective real-valued counterparts of these data are given, so we compared the estimated p-values based on the epistemic approach (see Table 4) with the "desirable" outputs (see Table 5). It seems that the p-values for the EKS/EcvM tests are very close to each other and to the results for the KS test (except *btp*) and not far from the p-value of the CvM test. The final decisions regarding the tested hypotheses, indicated by the methods proposed by us, are consistent with the conclusions of Hesamian and Taheri (2013).

To strengthen our reasoning, we added two "rather big" triangular values (namely, $(96, 100, 105)$ and

Table 4. Estimated p-values of tests on the street light equipment lifetimes.

Suppliers	ms-std	ms-anti	res-std	res-ant	avs-std	avs-anti	btp-std	btp-ant
EKS test								
A vs B	0.3185	0.2973	0.3182	0.2977	0.3176	0.2969	0.2385	0.2247
A vs C	0.1365	0.1263	0.1357	0.1265	0.1351	0.1263	0.0816	0.0753
ECvM test								
A vs B	0.3048	0.2968	0.3054	0.2971	0.3044	0.2965	0.2840	0.2758
A vs C	0.1090	0.1051	0.1093	0.1054	0.1083	0.1051	0.1038	0.1001

Table 5. Actual p-values of tests on the street light equipment lifetimes.

Suppliers	KS test	CvM test
A vs B	0.2857	0.2222
A vs C	0.1212	0.07879

$(104, 110, 117)$) to the data of supplier B and denoted this new set as supplier C. Once again, the p-values for the epistemic tests behave in a stable manner and are very close to the outputs of their classical counterparts (again, except *btp*).

The estimated p-values in the second example by Hesamian and Taheri (2013) concerning happiness of people were a little bigger (but still consistent) for the EKS/EcvM tests than their classical counterparts.

5.4.2. Electronic circuit thickness. In the work of Faraz and Shapiro (2010), fuzzy data concerning measurements of electronic circuit thickness are examined. This characteristic is essential during the production of electronic boards for vacuum cleaners. As noted, one subsample (number 21) makes the process out of control. Therefore, we checked the null hypothesis (2) when the first sample consists of the above-mentioned troublesome values, and the second one includes the rest of the observations. Then, the null hypothesis was not rejected (see Table 6) at the significance level 0.05, but the obtained p-values are significantly lower if they are compared under another data grouping (see Table 7). In this second analysis, the 21-st subsample was removed from the whole sample, and then the newly obtained set was divided into two samples (with 44 and 43 observations, respectively). Hence, it seems that the 21-st subsample is potentially troublesome, as indicated by Faraz and Shapiro (2010).

5.5. Comparison with other approaches. The works of Grzegorzewski (2020) or Grzegorzewski and Gadomska (2021), other approaches to goodness-of-fit tests for fuzzy data were presented. These methods are related to the ontic approach, which is different than the epistemic view for fuzzy data used in this paper. But, because of the lack of similar epistemic goodness-of-fit tests, we compare our method with the test by Grzegorzewski (2020) (further on abbreviated as *perm* test) as well as Grzegorzewski and Gadomska (2021) (*knn* test for short). The first one is an m-sample goodness-of-fit test, and its test statistic is based on the measure D_θ^λ proposed by Gil *et al*. (2002) and Trutschnig *et al*. (2009). Permutations of the initial data are used to approximate the respective p-value, so this procedure is called a "permutation test" by its authors. In the second test, which is also an m-sample goodness-of-fit test, the k-nearest neighbor approach is utilized for its test statistics together with the measure D_θ^λ and permutations of the original sample. For the *knn* test, we set $k = 5$ (i.e., five nearest neighbors, as advised by Grzegorzewski and Gadomska (2021)).

Firstly, as in Section 5.1, the shift in the location is analyzed. In Figs. 9 and 10, the estimated p-values are compared, when $X \sim \mathbb{F}_{(\mathrm{N,U,U})}$, and $K = 10, n = m = 10$. To facilitate understanding, only the most significant results (i.e., for *res-ant, avs-ant, btp-ant*) are reproduced from the previous figures. The p-values for the *perm* test seem to be lower (especially for the KS test) than for the respective KS or CvM test, and rather close to the *avs-ant* approach. This leads to a higher factor of H_0 rejection for the *perm* test, even when the second sample is without the shift. For the *knn* test, the respective p-values are lower and rather close to the ones for the KS test, but significantly higher for the CvM test. Then, the final answer concerning the null hypothesis can be more disturbed and dependent on the test applied. As for the power curves (Fig. 11), the *perm* method leads to very big values when compared with the KS test, and the *knn* approach—to low values for the CvM test. Therefore, their behavior seems to be more unstable than the case with the EKS/ECvM tests.

For the moderate sample ($n = m = 100$), the p-values are rather close to one another for larger values of the shift when $X \sim \mathbb{F}_{(\mathrm{N,U,U})}$. The *perm* test also leads to slightly lower p-values, even without the shift, which may be seen as a disadvantage of this method. On the contrary, the p-values for the *knn* test are significantly higher, especially for even moderate values of the shift (i.e. 0.25–0.75). This can lead to acceptance of the false null hypothesis.

Table 6. Estimated p-values of tests on the electronic circuit thickness (with the 21-st subsample).

Test	ms-std	ms-anti	res-std	res-ant	avs-std	avs-anti	btp-std	btp-ant
EKS	0.3185	0.2973	0.3182	0.2977	0.3176	0.2969	0.2385	0.2247
ECvM	0.1365	0.1263	0.1357	0.1265	0.1351	0.1263	0.0816	0.0753

Table 7. Estimated p-values of tests on the electronic circuit thickness (without the 21-st subsample).

Test	ms-std	ms-anti	res-std	res-ant	avs-std	avs-anti	btp-std	btp-ant
EKS	0.7899	0.8604	0.7910	0.8600	0.7909	0.8627	0.7845	0.8628
ECvM	0.7957	0.8739	0.7961	0.8729	0.7978	0.8742	0.7944	0.8737

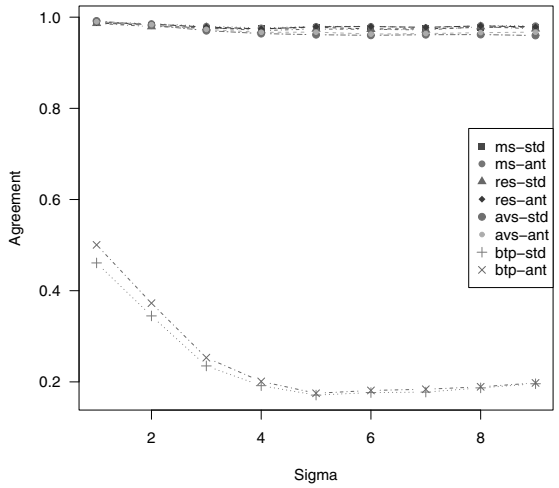

Fig. 8. Simulated ACC values for the KS test, $X \sim \mathbb{F}_{(\mathrm{N,U,U})}$, and $K = 10, n = m = 10$ (increasing σ).

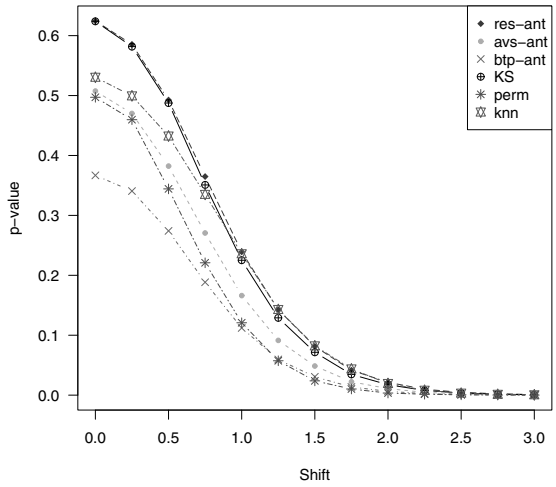

Fig. 9. Estimated p-values, $X \sim \mathbb{F}_{(\mathrm{N,U,U})}$, and $K = 10, n = m = 10$ (shift added to Y), averaged p-values: KS test based approaches.

Then, as in Section 5.2, the p-values for the shift in dispersion are compared. However, the *perm* test is not suitable for this case (see, e.g., Grzegorzewski and Gadomska, 2021). For the small initial sample (see Fig. 12), the p-values for the *knn* test are significantly lower (in the case of the KS test) or similar to the benchmark (for the CvM test), but with more unpredictable behavior than *avs-ant*—first, the p-values are bigger, then lower than the benchmark. The power curve for the *knn* test is significantly higher than its counterparts for the EKS/ECvM tests (which are close to our benchmarks; see, e.g., Fig. 13). For the moderate sample, the *knn* test leads to relatively higher p-values for the moderate change of σ for the second population (i.e., $\sigma \in [2, 3]$). This can be especially seen when the averaging of the p-values is applied.

For the street light equipment lifetimes (see Section 5.4.1), the p-values of both the *perm* and *knn* tests with 10000 permutations are very big—considerably higher than the outputs for their epistemic and classical counterparts (see Table 8). The same conclusions apply to the measurements of the electronic circuit thickness (see Section 5.4.2), where the obtained p-values do not indicate any problems with the troublesome 21st subsample (see Table 9), contrary to the case of Faraz and Shapiro (2010).

6. Conclusions

The epistemic bootstrap can be seen as a useful tool to support statistical inference for a certain type of imprecise data modeled by fuzzy sets. When the results of random experiments are somehow hidden, imprecisely defined, or measured, the epistemic perspective on fuzzy data perception and its analysis is required. However, as noted by Grzegorzewski and Romaniuk (2022b), the direct application of the extension principle in statistical inference to such data may be computationally unsatisfactory to potential users.

In this paper, we presented the epistemic bootstrap with several modifications as a universal basis for various statistical tests, involving their very important type—the goodness-of-fit tests. To check the quality of the proposed algorithms, we discussed two epistemic versions of the classical two-sample goodness-of-fit tests: the Kolmogorov–Smirnov and Cramer–von Mises tests.

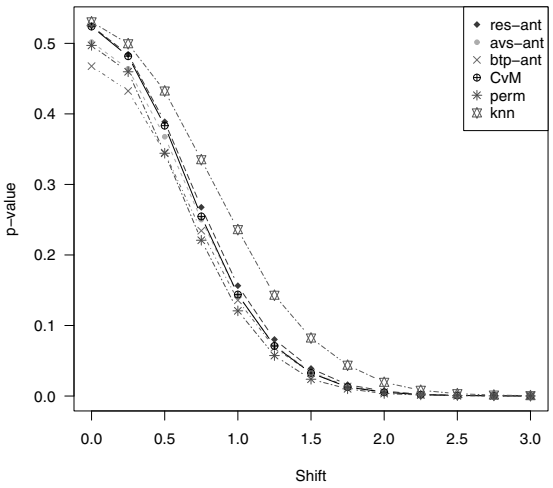

Fig. 10. Estimated p-values, $X \sim \mathbb{F}_{(\mathrm{N,U,U})}$, and $K = 10, n = m = 10$ (shift added to Y), averaged p-values: CvM test based approaches.

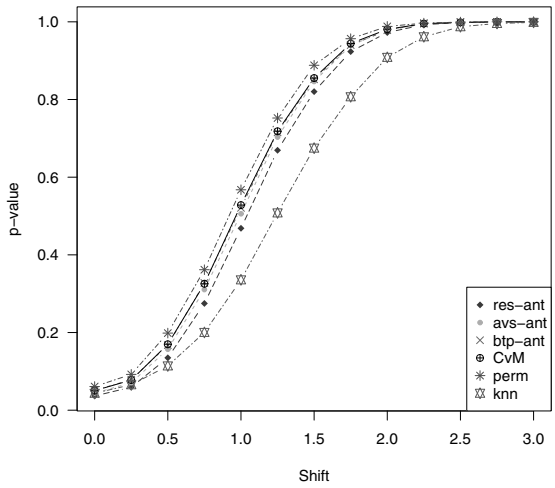

Fig. 11. Estimated power, $X \sim \mathbb{F}_{(\mathrm{N,U,U})}$, and $K = 10, n = m = 10$ (shift added to Y), averaged p-values: CvM test based approaches.

Table 8. Estimated p-values for other approaches: street light equipment.

Suppliers	*knn*	*perm*
A vs B	0.3442	0.3808
A vs C	0.2685	0.1488

Table 9. Estimated p-values for other approaches: electronic circuit thickness.

Subsample 21	*knn*	*perm*
Included	0.9389	0.4866
Removed	0.9711	0.8719

Based on both synthetic and real-life data, the epistemic versions of these tests were compared with their "crisp" counterparts that served as our benchmarks. It seems that the *avs-ant, ms-ant,* and *res-ant* approaches give the best results measured by both the similarity of the obtained p-values to the respective benchmarks and overall accuracy. Now these procedures are part of the R package *FuzzySimRes*. Moreover, the simple averaging of the p-values (instead of their combining by the Simes method) leads to an improvement in the results. The discussed epistemic tests were also compared with two other approaches known in the literature: the *perm* and *knn* tests. Also in this case our algorithms performed better and the obtained p-values seem to be more stable, outperforming inference based on these other tests.

Further research on resampling approaches for epistemic fuzzy data would be recommended. In particular, some hybrid methods, like combining the introduced epistemic tests with other nonparametric algorithms based on machine learning, e.g., the GAN (generative adversarial network), can lead to interesting results. These more sophisticated algorithms seem promising.

References

Anderson, T.W. (1962). On the distribution of the two-sample Cramer–von Mises criterion, *The Annals of Mathematical Statistics* **33**(3): 1148–1159.

Ban, A., Coroianu, L. and Grzegorzewski, P. (2015). *Fuzzy Numbers: Approximations, Ranking and Applications*, Polish Academy of Sciences, Warsaw.

Chernick, M.R., González-Manteiga, W., Crujeiras, R.M. and Barrios, E.B. (2011). Bootstrap methods, *in* M. Lovric (Ed.), *International Encyclopedia of Statistical Science*, Springer, Berlin/Heidelberg, pp. 169–174.

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

De Angelis, D. and Young, G.A. (1992). Smoothing the bootstrap, *International Statistical Review* **60**(1): 45–56.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife, *Annals of Statistics* **7**(1): 1–26.

Faraz, A. and Shapiro, A.F. (2010). An application of fuzzy random variables to control charts, *Fuzzy Sets and Systems* **161**(20): 2684–2694.

Gibbons, J.D. and Chakraborti, S. (2010). *Nonparametric Statistical Inference*, Chapman and Hall/CRC, New York.

Gil, M.A., Lubiano, M.A., Montenegro, M. and López, M.T. (2002). Least squares fitting of an affine function and strength of association for interval-valued data, *Metrika* **56**(2): 97–111.

Gil, M., Montenegro, M., González-Rodríguez, G., Colubi, A. and Casals, M. (2006). Bootstrap approach to the

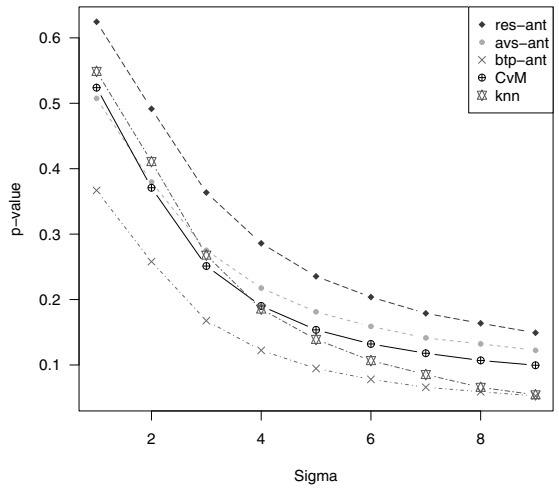

Fig. 12. Estimated p-values, $X \sim \mathbb{F}_{(\mathrm{N,U,U})}$, and $K = 10, n = m = 10$ (increasing σ), averaged p-values: CvM test based approaches.

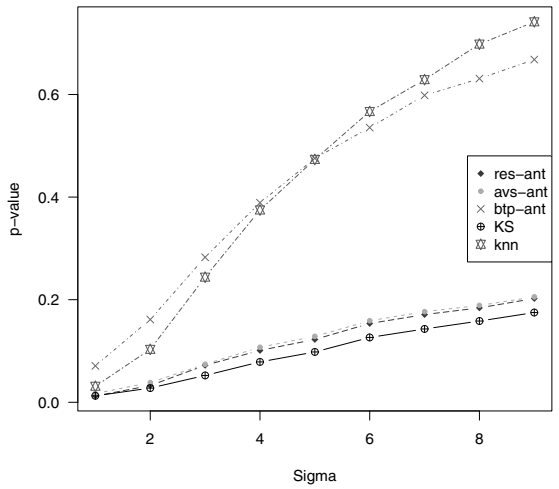

Fig. 13. Estimated power, $X \sim \mathbb{F}_{(\mathrm{N,U,U})}$, and $K = 10, n = m = 10$ (increasing σ), averaged p-values: KS test based approaches.

multi-sample test of means with imprecise data, *Computational Statistics and Data Analysis* **51**(1): 148–162.

González-Rodríguez, G., Montenegro, M., Colubi, A. and Gil, M. (2006). Bootstrap techniques and fuzzy random variables: Synergy in hypothesis testing with fuzzy data, *Fuzzy Sets and Systems* **157**(19): 2608–2613.

Grzegorzewski, P. (2008). Trapezoidal approximations of fuzzy numbers preserving the expected interval—Algorithms and properties, *Fuzzy Sets and Systems* **159**(11): 1354–1364.

Grzegorzewski, P. (2020). Permutation k-sample goodness-of-fit test for fuzzy data, *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, UK*, pp. 1–8.

Grzegorzewski, P. and Gadomska, O. (2021). Nearest neighbor tests for fuzzy data, *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Luxembourg*, pp. 1–6.

Grzegorzewski, P., Hryniewicz, O. and Romaniuk, M. (2019). Flexible bootstrap based on the canonical representation of fuzzy numbers, *Proceedings of the 11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019), Prague, Czech Republic*, pp. 490–497.

Grzegorzewski, P., Hryniewicz, O. and Romaniuk, M. (2020a). Flexible bootstrap for fuzzy data based on the canonical representation, *International Journal of Computational Intelligence Systems* **13**(1): 1650–1662.

Grzegorzewski, P., Hryniewicz, O. and Romaniuk, M. (2020b). Flexible resampling for fuzzy data, *International Journal of Applied Mathematics and Computer Science* **30**(2): 281–297, DOI: 10.34768/amcs-2020-0022.

Grzegorzewski, P. and Romaniuk, M. (2021). Epistemic bootstrap for fuzzy data, *Joint Proceedings of the IFSA-EUSFLAT-AGOP 2021 Conferences, Bratislavia, Slovakia*, pp. 538–545.

Grzegorzewski, P. and Romaniuk, M. (2022a). Bootstrap methods for epistemic fuzzy data, *International Journal of Applied Mathematics and Computer Science* **32**(2): 285–297, DOI: 10.34768/amcs-2022-0021.

Grzegorzewski, P. and Romaniuk, M. (2022b). Bootstrapped Kolmogorov–Smirnov test for epistemic fuzzy data, *in* D. Ciucci *et al.* (Eds), *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Springer International Publishing, Cham, pp. 494–507.

Hesamian, G., Akbari, M.G. and Shams, M. (2023). A goodness-of-fit test based on fuzzy random variables, *Fuzzy Information and Engineering* **15**(1): 55–68.

Hesamian, G. and Taheri, S. (2013). Linear rank tests for two-sample fuzzy data: A p-value approach, *Journal of Uncertain Systems* **7**(2): 129–137.

Kruse, R. (1982). The strong law of large numbers for fuzzy random variables, *Information Sciences* **28**(3): 233–241.

Kwakernaak, H. (1978). Fuzzy random variables. Part I: Definitions and theorems, *Information Sciences* **15**(1): 1–15.

Lubiano, M.A., Salas, A., Carleos, C., de la Rosa de Sáa, S. and Gil, M.A. (2017). Hypothesis testing-based comparative analysis between rating scales for intrinsically imprecise data, *International Journal of Approximate Reasoning* **88**: 128–147.

Lun, A. (2021). *metapod: Meta-Analyses on p-Values of Differential Analyses*, R package, `http://www.bioconductor.org/packages/release/bioc/html/metapod.html`.

Montenegro, M., Colubi, A., Casals, M. and Gil, M. (2004). Asymptotic and bootstrap techniques for testing the expected value of a fuzzy random variable, *Metrika* **59**: 31–49.

Romaniuk, M. and Grzegorzewski, P. (2023). Resampling fuzzy numbers with statistical applications: FuzzyResampling package, *The R Journal* **15**(1): 271–283.

Romaniuk, M., Grzegorzewski, P. and Parchami, A. (2023). *FuzzySimRes: Simulation and Resampling Methods for Epistemic Fuzzy Data*, R package, Version 0.2.0, https://CRAN.R-project.org/package=FuzzySimRes.

Romaniuk, M. and Hryniewicz, O. (2021). Discrete and smoothed resampling methods for interval-valued fuzzy numbers, *IEEE Transactions on Fuzzy Systems* **29**(3): 599–611.

Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance, *Biometrika* **73**(3): 751–754.

Smirnov, N. (1933). Estimate of deviation between empirical distribution functions in two independent samples, *Bulletin of Moscow University* **2**: 3–16.

Trutschnig, W., González-Rodríguez, G., Colubi, A. and Gil, M.A. (2009). A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread, *Information Sciences* **179**(23): 3964–3972.

Vovk, V. and Wang, R. (2020). Combining p-values via averaging, *Biometrika* **107**(4): 791–808.

Xiao, Y. (2012). *CvM2SL1Test: L1-Version of Cramer–von Mises Two Sample Tests*, R package, https://github.com/cran/CvM2SL1Test.

Przemysław Grzegorzewski holds an MSc in mathematics from the University of Warsaw as well as a PhD with distinction and a DSc (habilitation) in computer science from the Systems Research Institute of the Polish Academy of Sciences (SRI PAS). In 2018 he received his professorial title. He is currently a full professor at the Warsaw University of Technology and the SRI PAS. His areas of expertise include mathematical statistics, statistical decisions with imprecise data, data mining, fuzzy sets, fuzzy logic, soft computing, etc.

Maciej Romaniuk received his MSc degree in mathematics from the the University of Warsaw in 2001. He earned his PhD and DSc (habilitation) degrees in computer science from the Systems Research Institute of the Polish Academy of Sciences in 2007 and 2018, respectively. The main topics of his current research are Monte Carlo simulations, simulations of events under circumstances of uncertain and imprecise information, financial mathematics, actuarial mathematics, statistics, and fuzzy numbers.

Received: 26 July 2023
Revised: 16 November 2023
Re-revised: 12 January 2024
Accepted: 17 January 2024