Iranian Journal of Fuzzy Systems

Volume 16, Number 4, (2019), pp. 83-96

University of Sistan and Baluchestan

Asymptotic algorithm for computing the sample variance of interval data

A. Kołacz[1] and P. Grzegorzewski[2]

[1] *Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland*
[2] *Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland, and Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01-447 Warsaw, Poland*

A.Kolacz@mini.pw.edu.pl, pgrzeg@ibspan.waw.pl

Abstract

The problem of the sample variance computation for epistemic interval-valued data is, in general, NP-hard. Therefore, known efficient algorithms for computing variance require strong restrictions on admissible intervals like the no-subset property or heavy limitations on the number of possible intersections between intervals. A new asymptotic algorithm for computing the upper bound of the sample variance in a feasible time is proposed. Conditions required for its application with finite samples are discussed and some properties of the algorithm are also given. It appears that our new algorithm could be effectively applied in definitely more situations than methods used so far.

Keywords: Data analysis, interval data, sample variance.

1 Introduction

Big Data has become recently both a catchword as an important challenge for science and business. To define Big data in brief one may use just 3Vs: Volume, Velocity and Variety [21]. The first V correspond to the large amount of data that are generated by sensors, applications, etc. Velocity means that data are generated in a fast way and are expected to be processed rapidly. Finally, Variety denotes the complexity of data sets that consist of structured, unstructured or semi-structured data [29]. In particular, it may happen that the available information is delivered via interval-valued infogranules [25]. It happens quite often in engineering and industry, physics and environmental sciences, economy and social sciences, etc.

Looking at the problem of interval-valued data from the angle of statistics and machine learning one can list many interesting solutions in regression analysis [3, 24], time series [4], hypothesis testing [22], principal component analysis [8], clustering [7], classification and discriminant analysis [1, 9, 23, 26] (just to mention the most recent results).

However, traditional methods of the interval analysis may not be sufficient if the data sets are too large. Indeed, many computational problems connected with interval-valued data are NP-hard, which means, roughly speaking, that in general no computationally efficient algorithm can solve all particular cases of the problem under study [16]. All in all even so specific Big Data problem like interval-valued data bring new challenges and demands new approaches and algorithms that might be helpful to make decisions.

At the very beginning of any discussion on interval-valued data one should become aware that a closed interval may be used to model the following two basic types of information: the imprecise description of a point-valued quantity (epistemic set) or the precise description of a set-valued entity (ontic set). An *epistemic* (disjunctive) set X contains an ill-known actual value of a point-valued quantity x, so we can write $x \in X$. It represents the epistemic state of an agent, hence it does not exist per se (philosophically, it is *de dicto*). Sets that represent collections of elements forming composite objects are called *ontic* (conjunctive). A conjunctive set is the precise representation of an objective entity (philosophically, it is a *de re notion*). An ontic set X is the value of a set-valued variable V, so we can write $V = X$ (for more details see [5]).

Corresponding Author: P. Grzegorzewski
Received: January 2018; Revised: December 2018; Accepted: December 2018.

Further on we restrict our attention to data analysis based on intervals perceived from the epistemic perspective only. The monograph [20] by Nguyen et al. could be recommended as an excellent guide to statistics under interval uncertainty in this approach. Although many results for interval data are obtained as straightforward generalizations of the corresponding situations known for real data, one might be surprised that sometimes serious difficulties appear even in problems which at first glance seem elementary. One of such cases is the sample variance computation for interval data which may cause problems even for a very small sample. What is worse, the problem of computing variance is, in general, NP-hard [14]. Therefore, known efficient algorithms for computing variance require heavy restrictions on admissible intervals. This is the reason why an algorithm which could be applied effectively in more practical cases is still of interest (for some newest approaches see, e.g. [15, 17, 16, 20]). Similar problem appears in the problem of estimating correlation under interval uncertainty [12].

In this paper we propose a new asymptotic algorithm for computing the upper bound of the sample variance. As it is known, the lower bound of the sample variance can be always computed in a feasible (polynomial) time. Therefore, the suggested approach solves the problem of the sample variance estimation in many situations. Moreover, it appears that our new algorithm could be effectively applied in much wider class of interval-valued data sets than methods available so far.

Although our considerations in this paper are restricted to interval-valued data, it is worth noting that the proposed algorithm might be also useful for calculating an upper bound of the sample variance of fuzzy data. Indeed, according to the epistemic perspective, a fuzzy random sample may be treated as a fuzzy perception of the usual real-valued random sample. Fuzzy observations in statistics are usually modeled by fuzzy numbers, i.e. such fuzzy sets which are normal, fuzzy-convex, which have a bounded support and upper semicontinuous membership function. Basing on the resolution identity, each fuzzy number can be described equivalently as a monotonic family of closed intervals, so-called α-cuts. Further on, all calculations on fuzzy numbers can be performed using intervals. Consequently, one can estimate the upper bound of a fuzzy sample variance applying the suggested algorithm dedicated to interval data.

The paper is organized as follows. In Section 2 we introduce basic notation related to interval-valued data. In Section 3 we demonstrate some practical problems that may occur in the sample variance estimation for epistemic data and we instruct how to cope with them. We also discuss limitations in the variance computations for this type of data. Next, in Section 4 we prove a few lemmas that form a framework for our algorithm which is presented in Section 5. Conditions required for application of the proposed new algorithm to finite samples are also discussed there. Finally, some results of the simulation study performed to examine the properties of the suggested algorithm and to compare it with other methods are given in Section 6.

2 Basic notation

Let $\mathcal{K}_c(\mathbb{R}) = \{[u, v] : u, v \in \mathbb{R},\ u < v\}$ denote the family of all non-empty closed and bounded intervals on the real line $\mathbb{R}$. Each compact interval $X \in \mathcal{K}_c(\mathbb{R})$ can be expressed by its endpoints, i.e. $X = [\underline{x}, \overline{x}]$. Alternatively, we may consider the notation $X = [\operatorname{mid} X \pm \operatorname{spr} X]$, with $\operatorname{spr} X > 0$, where $\operatorname{mid} X = \frac{1}{2}(\underline{x} + \overline{x})$ is the mid-point (center) of the interval X, while $\operatorname{spr} X = \frac{1}{2}(\overline{x} - \underline{x})$ is the spread (radius) of X.

The Minkowski addition and the product by scalars define a natural arithmetic on $\mathcal{K}_c(\mathbb{R})$, i.e. for any $X, Y \in \mathcal{K}_c(\mathbb{R})$ and $\lambda \in \mathbb{R}$, $X + Y = \{x + y : x \in X, y \in Y\}$, and $\lambda X = \{\lambda x : x \in X\}$. Thus, using the inf / sup-representation, basic operations on intervals $X = [\underline{x}, \overline{x}]$ and $Y = [\underline{y}, \overline{y}]$ are given by

$$X + Y = [\underline{x} + \underline{y}, \overline{x} + \overline{y}], \quad X - Y = [\underline{x} - \overline{y}, \overline{x} - \underline{y}], \text{ and } \quad \lambda X = [\min\{\lambda\underline{x}, \lambda\overline{y}\}, \max\{\lambda\underline{x}, \lambda\overline{y}\}].$$

These two operations can be jointly expressed in terms of the mid /spr -representation as follows

$$X + \lambda Y = [(\operatorname{mid} X + \lambda \operatorname{mid} Y) \pm (\operatorname{spr} X + |\lambda| \operatorname{spr} Y)].$$

It is worth noting that $(\mathcal{K}_c(\mathbb{R}), +, \cdot)$ is not linear but semilinear: in general, $X + (-1)X \neq \{0\}$, unless $X = \{x\}$. Moreover, the Minkowski difference does not satisfy, in general, the addition/subtraction property $(X + (-1)Y) + Y = X$.

The multiplication of two intervals $X = [\underline{x}, \overline{x}]$ and $Y = [\underline{y}, \overline{y}]$ is given by

$$X \cdot Y = [\min\{\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y}\}, \max\{\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y}\}], \tag{1}$$

while their division is defined as follows $X/Y = [\min\{\underline{x}/\underline{y}, \underline{x}/\overline{y}, \overline{x}/\underline{y}, \overline{x}/\overline{y}\}, \max\{\underline{x}/\underline{y}, \underline{x}/\overline{y}, \overline{x}/\underline{y}, \overline{x}/\overline{y}\}]$, provided $0 \notin [\underline{y}, \overline{y}]$. Interval computations in their current form were independently invented by Warmus [30], Sunaga [27] and Moore [18].

Suppose a sequence $X_1 = [\underline{x}_1, \overline{x}_1], \ldots, X_n = [\underline{x}_n, \overline{x}_n]$ denotes interval perceptions of unknown true outcomes $x_1, \ldots, x_n$ of the experiment, where $x_i \in X_i$. Let us consider a function $f(x_1, \ldots, x_n)$. Our goal is to find its range for all possible outcomes of the experiment, i.e. $W = W(X_1, \ldots, X_n) = \{f(x_1, \ldots, x_n) : x_1 \in X_1, \ldots, x_n \in X_n\}$.

If f is continuous its range is also an interval $[\underline{w}, \overline{w}]$. Hence, to obtain W it is sufficient to find $\underline{w}$ and $\overline{w}$. For example, since the arithmetic mean $m(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$ is monotonically increasing and continuous for each argument, the range $M = [\underline{m}, \overline{m}]$ of the mean for interval data $X_1 = [\underline{x}_1, \overline{x}_1], \ldots, X_n = [\underline{x}_n, \overline{x}_n]$ is given by

$$M = \Big[\frac{1}{n}\sum_{i=1}^{n} \underline{x}_i, \frac{1}{n}\sum_{i=1}^{n} \overline{x}_i\Big]. \tag{2}$$

Unfortunately, it is not always easy or even possible to find the actual range of W. This is why we usually try to compute at least the interval $\widetilde{W}$, such that $\widetilde{W} \supseteq W$, called an *enclosure* for W. If $\widetilde{W} = W$ we say that the enclosure is *exact* (e.g. (2) gives the exact enclosure of the arithmetic mean for the interval data). Unluckily, the straightforward application of the formulas for the interval addition, subtraction, multiplication and division discussed above usually leads to overestimation of the exact range. Moreover, it may also happen that finding a desired enclosure is not feasible. We will discuss this in more details in the subsequent sections.

3 Computing variance - general problems

Let us start with a simple example illustrating some possible sources of problems that may appear in computations with interval data. Consider the well-known sample variance $s^2(x_1, \ldots, x_n) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2$, where m stands for the mean. Theoretically, computation of the sample variance should not cause any problem since it is continuous, so one may expect to easily obtain its range $S^2 = \left[\underline{S^2}, \overline{S^2}\right]$ for $[\underline{x}_1, \overline{x}_1] \times \ldots \times [\underline{x}_n, \overline{x}_n]$. Unfortunately, one may face some traps and fundamental problems just from the very beginning of calculations.

Example 3.1. *Suppose we have three observations described by the following intervals: $X_1 = [0, 1]$, $X_2 = [0, 1]$ and $X_3 = [0, 1]$. If we look at them from the ontic perspective, the variance of this sample is zero – indeed, since we have three identical objects there is no variation of outcomes and consequently any reasonable measure of dispersion of such a sample would indicate zero (see [11, 13] for some general properties of measures of dispersion). However, from the epistemic perspective discussed in this paper, these three intervals, although looking identically, are only unprecise perceptions of some unknown real outcomes of an experiment and they can vary vastly. Therefore, it would be desirable to find a range of the sample variance corresponding to all possible values of the experimental results described by those intervals.*

Someone who tries to apply directly the rules of interval arithmetic described in Section 2 to obtain $S^2(X_1, X_2, X_3)$ might be surprised since, by such straightforward calculation, they get $M = [0, 1]$ and $X_i - M = [-1, 1]$, for $i = 1, 2, 3$, which yields $S^2 = [-1.5, 1.5]$. Obviously this is a strange result because the variance by the definition cannot be negative. Thus, one may ask immediately: Why did it happen so? And how to improve calculations so they would produce reasonable results?

One source of problems met in the aforementioned calculations is connected with the so-called interval dependency phenomenon. Roughly speaking, it means that the manner of choosing points from the intervals must be finer in the case of nonmonotone functions. Actually, a direct computation of $X_i - M$ is incorrect since X_i is one of the summands in the expression for M. Taking this into account we may rewrite this expression as follows

$$X_i - M = \Big(1 - \frac{1}{n}\Big)X_i - \frac{1}{n}\sum_{j \neq i} X_i. \tag{3}$$

Now substituting $X_i = [0, 1]$, $i = 1, 2, 3$, into (3) we get much narrower interval $X_i - M = \frac{2}{3}[0, 1] - \frac{1}{3}[0, 2] = \left[-\frac{2}{3}, \frac{2}{3}\right]$ than $[-1, 1]$ obtained previously.

Next, keeping in mind the interval dependency, one should notice that for intervals $X^2 \neq X \cdot X$. Indeed, $X^2 = \{x^2 : x \in X\}$ while the square calculated as a product according to (1) actually gives $X \cdot X = \{xy : x, y \in X\}$. Therefore, the square required in the sample variance should be calculated as follows (see [19])

$$X^2 = \begin{cases} [0, \max\{\underline{x}^2, \overline{x}^2\}] & \text{if} \quad 0 \in [\underline{x}, \overline{x}] \\ [\min\{\underline{x}^2, \overline{x}^2\}, \max\{\underline{x}^2, \overline{x}^2\}] & \text{if} \quad 0 \notin [\underline{x}, \overline{x}]. \end{cases} \tag{4}$$

By (4) we obtain $(X_i - M)^2 = \left[0, \frac{4}{9}\right]$ *whereas* $(X - M) \cdot (X - M) = [-1, 1]$. *Hence, calculating appropriately the squares of the differences between successive observations and their average we obtain* $S^2 = \frac{1}{2}\sum_{i=1}^{3}\left[0, \frac{4}{9}\right] = \left[0, \frac{2}{3}\right]$.

Although we have gotten rid of the negative values, the last enclosure of the sample variance is not exact, because there is still some excess in width. To improve the result we should consider carefully in which operation not yet discussed there is still some interval dependency. One may notice that it is not reasonable to sum up all the possible values of $(x_i - m)^2$, *where* $x_i \in X_i$ *and* $m \in M$ *because each choice of* $x_1, \ldots, x_n$ *determines a single value of* $m = \frac{1}{n}\sum_{i=1}^{n} x_i$. *Thus we have to incorporate this remark into our computations.*

Let us adopt the following notation: $a_i = x_i - m$. *Obviously* $a_1 + a_2 + a_3 = 0$. *Therefore, our goal is to find the minima and maxima of the function* $F(a_1, a_2, a_3) = a_1^2 + a_2^2 + a_3^2$ *subject to* $a_1 + a_2 + a_3 = 0$, *where each* $a_i \in \left[-\frac{2}{3}, \frac{2}{3}\right]$, *as it is shown above. Hence, applying the method of Lagrange multipliers we consider the Lagrangian* $\mathcal{L}(a_1, a_2, a_3, \lambda) = a_1^2 + a_2^2 + a_3^2 - \lambda(a_1 + a_2 + a_3)$, *where* λ *is the Lagrange multiplier. Some easy calculations yields* $\min F(a_1, a_2, a_3) = F(0, 0, 0) = 0$. *To find the desired maximum we have to consider the border of the domain of our function. Because of the symmetry we get immediately that maximum is reached if any* $a_i = \frac{2}{3}$ *or* $a_i = -\frac{2}{3}$ *and two other arguments are identical and equal either to* $-\frac{1}{3}$ *or to* $\frac{1}{3}$. *At each such point* $\max F = \frac{2}{3}$. *Now, since* $s^2 = \frac{1}{2}F$ *we obtain the desired range of the sample variance, i.e.* $S^2 = \left[0, \frac{1}{3}\right]$ *and this is the exact enclosure of the sample variance for the given data set.*

Please, note that a similar example illustrating problems that appear using straightforward interval computations can be found in [20]. There one can also find some advices how to decrease excess in estimating desired enclosures (e.g. the bisection method, the Taylor expansion).

Although we have shown that a careful study of interval dependencies may help in the sample variance computation it does not mean that the problem has always a satisfactory solution. Unfortunately, optimization methods sometimes may last too long, especially if n is large. It can be shown that the lower endpoint $\underline{S^2}$ can be always computed in a feasible (polynomial) time. However, the problem of computing the upper endpoint $\overline{S^2}$ is, in general, NP-hard (see [20]). It is worth noting that this very fact was firstly proven by Vavasis [28].

Since in general we cannot efficiently compute the upper endpoint $\overline{S^2}$ of the sample variance, it would be desirable to distinguish situations when such computation is possible. The following three cases of interval data sets for which an efficient algorithm to compute $\overline{S^2}$ exists are mentioned in [6, 10, 20, 31]:

- intervals satisfy the *no-subset property*, i.e. if for every two different intervals $[\underline{x}_i, \overline{x}_i]$ and $[\underline{x}_j, \overline{x}_j]$ we have $[\underline{x}_i, \overline{x}_i] \nsubseteq (\underline{x}_j, \overline{x}_j)$,
- intervals can be divided into m subclasses and within each of these classes the no-subset property is satisfied,
- when for some integer $c_0 \geqslant 2$ every group of c_0 intervals has an empty intersection.

It can be shown that for interval-valued data belonging to these three classes there exist $O(n \cdot \log(n))$, $O(n^m)$ and $O(n \cdot \log(n))$ time algorithms for computing $\overline{S^2}$, respectively (these algorithms as well as their justifications can be found in [20]).

Recently an heuristic method for computing variance based on a genetic algorithm and performing on a grid computing infrastructure was presented in [2].

Further on we propose a new algorithm based on some asymptotic reasoning that enables to estimate $\overline{S^2}$ effectively for a broader family of interval data sets than those three aforementioned classes.

4 Some asymptotic results

It has been shown in [20] that to get the maximum value of the sample variance it is sufficient to find the largest value of $S^2(v_1, \ldots, n_n)$ for all 2^n possible combinations $v_1, \ldots, v_n$, where $v_i \in \{\underline{x}_i, \overline{x}_i\}$, $i = 1, \ldots, n$. Although such calculations can be troublesome if n is large, just the idea indicating where to choose points maximizing the sample variance is worth remembering.

Remark 4.1. *There exists a sequence* $x_1^*, \ldots, x_n^*$, *where* $x_i^* \in \{\underline{x}_i, \overline{x}_i\}$, *for which the sample variance attains its upper bound, i.e.* $\overline{S^2} = S^2(x_1^*, \ldots, x_n^*) = \max\{S^2(v_1, \ldots, v_n) : v_i \in \{\underline{x}_i, \overline{x}_i\}, i = 1, \ldots, n\}$.

Now the central problem is to discover a method of identifying $x_1^*, \ldots, x_n^*$ without the need of calculating and comparing 2^n possible combinations of candidates for the maximizing endpoints. Before we propose such a method let us consider some auxiliary lemmas which appear helpful in characterizing $x_1^*, \ldots, x_n^*$. Further on the following notation would be also useful: $x_L = \max\{x_i^* : x_i^* = \underline{x}_i\}$, and $x_R = \min\{x_i^* : x_i^* = \overline{x}_i\}$.

Lemma 4.2. *Let $x_1^*, \ldots, x_n^*$, where $x_i^* \in \{\underline{x}_i, \overline{x}_i\}$, $i = 1, \ldots, n$, denote a sequence such that $\overline{S^2} = S^2(x_1^*, \ldots, x_n^*)$. Then $x_L < x_R$.*

Proof. Assume, on the contrary, that $x_L \geqslant x_R$. Moreover, denote the intervals containing x_L and x_R by X_L and X_R, respectively. Obviously, $x_L < \operatorname{mid} X_L$ and $\operatorname{mid} X_R < x_R$.

Let us recall the following iterative formula for computing the sample variance

$$S^2 = \frac{n-2}{n-1} S_{-i}^2 + \frac{1}{n}(x_i - m_{-i})^2, \tag{5}$$

where $S^2 = S^2(x_1, \ldots, x_n)$, while S_{-i}^2 and m_{-i} denote the sample variance and the arithmetic mean, respectively, calculated for all observations except of x_i, i.e. $S_{-i}^2 = S^2(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ and $m_{-i} = \frac{1}{n-1}(x_1 + \ldots + x_{i-1} + x_{i+1} + \ldots + x_n)$. Without loss of generality let us assume that $X_R = X_i$. Therefore, since the sequence $x_1^*, \ldots, x_n^*$ maximizes the sample variance, we have

$$S^2(x_1^*, \ldots, x_{i-1}^*, \overline{x}_i, x_{i+1}^*, \ldots, x_n^*) > S^2(x_1^*, \ldots, x_{i-1}^*, \underline{x}_i, x_{i+1}^*, \ldots, x_n^*). \tag{6}$$

Now, applying (5) to both sides of inequality (6) we obtain $(\overline{x}_i - m_{-i}^*)^2 > (\underline{x}_i - m_{-i}^*)^2$, where $m_{-i}^* = \frac{1}{n-1}\sum_{j \neq i} x_j^*$ and after some simple transformations we obtain $\frac{1}{2}(\overline{x}_i + \underline{x}_i) > m_{-i}^*$. This result can be written as $\operatorname{mid} X_i > m_{-i}^*$ which means that $\operatorname{mid} X_R > m_{-R}^*$.

Similarly, assuming without loss of generality that $X_L = X_j$, we have

$$S^2(x_1^*, \ldots, x_{j-1}^*, \underline{x}_j, x_{j+1}^*, \ldots, x_n^*) > S^2(x_1^*, \ldots, x_{j-1}^*, \overline{x}_j, x_{j+1}^*, \ldots, x_n^*),$$

which leads to $\operatorname{mid} X_L < m_{-L}^*$.

Now, combining these results with the properties of X_L and X_R stated above, we conclude that $x_R > m_{-R}^*$ and $x_L < m_{-L}^*$ and hence, by the assumptions that $x_L \geqslant x_R$, we obtain $m_{-R}^* < m_{-L}^*$. On the other hand if $x_L \geqslant x_R$, then obviously $m_{-R}^* \geqslant m_{-L}^*$ by the properties of the average. Thus we have obtained a contradiction, and the lemma is proved. □

Remark 4.3. *Notice that equation* (4) *implies the following general result: if for some i we fix $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$, then the sample variance is a quadratic function of x_i. The minimum of this function is at m_{-i}, so if we are given a choice of x_i to maximize the variance, e.g. $x_i \in \{a, b\}$, then we need to choose a point whose distance from m_{-i} is the largest.*

Keeping in mind that the maximum value of the sample variance is attained at some endpoints of the intervals, i.e. for some sequence of points $v_1, \ldots, v_n$, where $v_i \in \{\underline{x}_i, \overline{x}_i\}$, $i = 1, \ldots, n$, it would be sometimes convenient to rewrite these endpoint using mid /spr notation as follows: $v_i = \operatorname{mid} X_i + (-1)^{\varepsilon_i} \operatorname{spr} X_i$, where $\varepsilon_i = \begin{cases} 0 & \text{if} \quad v_i = \overline{x}_i, \\ 1 & \text{if} \quad v_i = \underline{x}_i. \end{cases}$

Moreover, let

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \operatorname{mid} X_i \tag{7}$$

denote the mean of the centers of all intervals and let

$$\delta(\varepsilon_1, \ldots, \varepsilon_n) = \frac{1}{n}\sum_{i=1}^{n} (-1)^{\varepsilon_i} \operatorname{spr} X_i. \tag{8}$$

Theorem 4.4. *Let $x_1^*, \ldots, x_n^*$ denote the endpoints of intervals $X_1, \ldots, X_n$ which maximize the sample variance and whose centers and spreads form bounded sequences. Then, assuming that both limits exist, the following equality holds*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \operatorname{mid} X_i = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i^*. \tag{9}$$

Proof. Denoting the mean of variance maximizers by $m^* = \frac{1}{n}\sum_{i=1}^{n} x_i^*$ our thesis (9) states that μ and m^* approach the same limit as $n \to \infty$. We will prove this statement in a few steps.

(i) Firstly, let us consider the relationship between the mean of all variance maximizers m^* and the mean of all but one maximizers, where x_i^* has been eliminated, i.e. m_{-i}^*. One can easily see that $m_{-i}^* = \frac{n}{n-1}m^* - \frac{1}{n-1}x_i^*$. Hence, by a straightforward transformation we obtain

$$x_i^* - m_{-i}^* = \frac{n}{n-1}(x_i^* - m^*), \tag{10}$$

which implies that for each i the mean m_{-i}^* has the same limit as m^* when n tends to infinity.
(ii) Now let us assume that we know the mean of $n-1$ variance maximizers m_{-i}^* and our goal is to choose such a point from interval X_i that the sample variance based on n points would reach its upper bound. By (4) we conclude that to maximize the sample variance for some fixed $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ we have to choose a point $x_i \in X_i$ which is the most distant from m_{-i}^*. Therefore, if $m_{-i}^* < \operatorname{mid} X_i$, then $x_i^* = \overline{x}_i$, while if $\operatorname{mid} X_i < m_{-i}^*$, then $x_i^* = \underline{x}_i$. It means that m_{-i}^* determines the choice of the i-th maximizing endpoint for all $i = 1, \ldots, n$.
(iii) Now let us start from a general remark. Given are intervals $A_1 = [a_1, b_1], ...,$ $A_n = [a_n, b_n]$ and any $n_a, n_b \geqslant 1$ such that $n_a + n_b = n$. We choose arbitrarily n_a left endpoints and from the remaining n_b intervals we choose right endpoints. Without loss of generality we can denote them by $a_1, \ldots, a_{n_a}$ and $b_{n_a+1}, \ldots, b_{n_b}$. Computing the mean of them we get $\overline{A} = \frac{\sum_{i=1}^{n_a} a_i + \sum_{i=n_a+1}^{n_b} b_i}{n}$ and the mean of the endpoints that were not used $\overline{B} = \frac{\sum_{i=1}^{n_a} b_i + \sum_{i=n_a+1}^{n_b} a_i}{n}$.

Consider the following sequence

$$y_1 = a_1 + (b_1 - a_1)y, \ldots, y_{n_a} = a_{n_a} + (b_{n_a} - a_{n_a})y, y_{n_a+1} = b_{n_a+1} - (b_{n_a+1} - a_{n_a+1})y, \ldots, y_{n_b} = b_{n_b} - (b_{n_b} - a_{n_b})y,$$

where $y \in [0, 1]$. It is clear that for every $y \in [0, 1]$, the mean $\frac{1}{n}\sum_{i=1}^{n} y_i$ belongs to the interval $[\overline{A}, \overline{B}]$. Indeed

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{\sum_{i=1}^{n_a} a_i + \sum_{i=n_a+1}^{n_b} b_i}{n} + \frac{1}{n}\left(y\sum_{i=1}^{n}(b_i - a_i)\right) = \overline{A} + \frac{1}{n}\left(y\sum_{i=1}^{n}(b_i - a_i)\right)$$

and we see that $\frac{1}{n}\sum_{i=1}^{n} y_i = \overline{A}$ for $y = 0$ and $\frac{1}{n}\sum_{i=1}^{n} y_i = \overline{B}$ for $y = 1$. Given that the mean is a monotone function of its arguments the remark is proven. In particular, $y = 0.5$ yields that the mean of the interval centers belongs to $[\overline{A}, \overline{B}]$.

Let us divide the centers of all intervals, i.e. $midX_1, ..., midX_n$, into two subsets J_L and J_R such that

$$J_L = \{\operatorname{mid} X_i : x_i^* = \underline{x}_i\}, \quad \text{and} \quad J_R = \{\operatorname{mid} X_i : x_i^* = \overline{x}_i\}.$$

Moreover, let us arrange the centers $\operatorname{mid} X_1, \ldots, \operatorname{mid} X_n$ into the nondecreasing order $t_1 \leqslant t_2 \leqslant \ldots \leqslant t_n$, where t_i denotes the i-th greatest center. Without loss of generality let us assume that $t_k = \max\{t_i : t_i \in J_L\}$ and $t_{k+1} = \min\{t_i : t_i \in J_R\}$.

By the facts stated in (i), (ii) and Lemma 4.2 we may conclude that $t_k \leqslant t_{k+1}$ (equality holds, for example, if a sample consists of n copies of the same interval).

Recall that the variance maximizing endpoints were denoted by $x_1^*, \ldots, x_n^*$ and that their mean $m^* \in [t_k, t_{k+1}]$ which follows from (i) and (ii). Let us now consider the endpoints opposite to $x_1^*, \ldots, x_n^*$. Denote them by $x_1^\circ, \ldots, x_n^\circ$, i.e. x_i° is the opposite endpoint to x_i^* for all $i = 1, \ldots, n$. Also let $m^\circ = \frac{1}{n}\sum_{i=1}^{n} x_i^\circ$ and $m_{-i}^\circ = \frac{1}{n-1}\sum_{j \neq i} x_j^\circ$. and x_k°, x_{k+1}° denote the non-maximizing endpoints of intervals whose centers are t_k and t_{k+1} respectively. There are four possible scenarios:

1. $m_{-k}^\circ > m^\circ$ and $m_{-(k+1)}^\circ > m^\circ$,
2. $m_{-k}^\circ \leqslant m^\circ$ and $m_{-(k+1)}^\circ \leqslant m^\circ$,
3. $m_{-k}^\circ \leqslant m^\circ$ and $m_{-(k+1)}^\circ \geqslant m^\circ$,
4. $m_{-k}^\circ > m^\circ$ and $m_{-(k+1)}^\circ \leqslant m^\circ$.

Following point (ii) we conclude that

1. $m_{-k}^\circ > \operatorname{mid} X_k$ and $m_{-(k+1)}^\circ < \operatorname{mid} X_{k+1}$,

2. $m^{\circ}_{-k} \leqslant \operatorname{mid} X_k$ and $m^{\circ}_{-(k+1)} \geqslant \operatorname{mid} X_{k+1}$,
3. $m^{\circ}_{-k} \leqslant \operatorname{mid} X_k$ and $m^{\circ}_{-(k+1)} \leqslant \operatorname{mid} X_{k+1}$,
4. $m^{\circ}_{-k} > \operatorname{mid} X_k$ and $m^{\circ}_{-(k+1)} \geqslant \operatorname{mid} X_{k+1}$.

And taking (i) into account, which means that as $n \to \infty$, $m^{\circ}_{-(k+1)} = m^{\circ}_{-k} = m^{\circ}$, we have

1. $m^{\circ} \in (\operatorname{mid} X_k, \operatorname{mid} X_{k+1})$,
2. $m^{\circ} = \operatorname{mid} X_k = \operatorname{mid} X_{k+1}$,
3. $m^{\circ} = \operatorname{mid} X_{k+1}$,
4. $m^{\circ} = \operatorname{mid} X_k$.

We showed that in general $m^{\circ} \in [\operatorname{mid} X_k, \operatorname{mid} X_{k+1}]$ and $m^{*} \in [\operatorname{mid} X_k, \operatorname{mid} X_{k+1}]$. But if we now apply the initial remark from point (iii) to $X_1, \ldots, X_n$ and choose the left and right maximizing endpoints as sequences (a_i) and (b_i) respectively, then we conclude that $\mu = \frac{1}{n}\sum_{i=1}^{n} \operatorname{mid} X_i \in [\operatorname{mid} X_k, \operatorname{mid} X_{k+1}]$.

(iv) Now adopting mid /spr notation for $v_i \in \{\underline{x}_i, \overline{x}_i\}$ and keeping in mind the notation given in (7) and (8) we have

$$\begin{aligned}
(n-1)S^2(v_1,\ldots,v_n) &= \sum_{i=1}^{n}(v_i - \frac{1}{n}\sum_{i=1}^{n} v_i)^2 = \sum_{i=1}^{n}\big(\operatorname{mid} X_i + (-1)^{\varepsilon_i}\operatorname{spr} X_i - \mu - \delta(\varepsilon_1,\ldots,\varepsilon_n)\big)^2 \\
&= \sum_{i=1}^{n}(\operatorname{mid} X_i - \mu)^2 + 2\sum_{i=1}^{n}\big((-1)^{\varepsilon_i}\operatorname{spr} X_i - \delta(\varepsilon_1,\ldots,\varepsilon_n)\big)(\operatorname{mid} X_i - \mu) \\
&+ \sum_{i=1}^{n}\big((-1)^{\varepsilon_i}\operatorname{spr} X_i - \delta(\varepsilon_1,\ldots,\varepsilon_n)\big)^2 \\
&= \sum_{i=1}^{n}(\operatorname{mid} X_i - \mu)^2 + 2\sum_{i=1}^{n}(-1)^{\varepsilon_i}(\operatorname{mid} X_i - \mu)\operatorname{spr} X_i + \sum_{i=1}^{n}(\operatorname{spr} X_i)^2 - n\delta^2(\varepsilon_1,\ldots,\varepsilon_n). \qquad (11)
\end{aligned}$$

By the results given in (iii) we may conclude that if n is large enough, then $\operatorname{sgn}(\operatorname{mid} X_i - \mu) = (-1)^{\varepsilon_i}$ or $\operatorname{mid} X_i - \mu = 0$. Hence maximization of the sample variance is equivalent to the minimization of $|\delta(\varepsilon_1,\ldots,\varepsilon_n)|$ which is the only element in (11) that depends on the choice of the interval endpoints.

(v) By (8) the minimization of $|\delta(\varepsilon_1,\ldots,\varepsilon_n)|$ is equivalent to the minimization of $|\sum_{i=1}^{n}(-1)^{\varepsilon_i}\operatorname{spr} X_i|$. In other words, we want to assign either -1 or 1 to the consecutive interval spreads so that $|\sum_{i=1}^{n}(-1)^{\varepsilon_i}\operatorname{spr} X_i|$ is as close to zero as possible.

Let us now consider an interval $I = (0, \sum_{i=1}^{n}\operatorname{spr} X_i)$. It is obvious that it can be given as a union of n intervals $I_1, \ldots, I_n$, such that the length of I_j equals $\operatorname{spr} X_j$ for $j = 1, \ldots, n$, i.e. $I = \bigcup_{j=1}^{n} I_j$. Obviously, it holds also for any permutation π of those intervals, i.e. $I = \bigcup_{j=1}^{n} I_{\pi(j)}$.

Let c denote the center of I. For every permutation of the intervals $I_1, \ldots, I_n$ one of the following situations is possible: either c lies at the left endpoint of I_k or c lies within I_k for some k. Now we assign -1 to $\operatorname{spr} X_1, \ldots, \operatorname{spr} X_{k-1}$ and 1 to $\operatorname{spr} X_k, \ldots, \operatorname{spr} X_n$. If c lies at the left endpoint of I_k then $|\sum_{i=1}^{n}(-1)^{\varepsilon_i}\operatorname{spr} X_i| = 0$, otherwise $|\sum_{i=1}^{n}(-1)^{\varepsilon_i}\operatorname{spr} X_i| < \operatorname{spr} X_k$. Hence, the actual minimal value of $|\sum_{i=1}^{n}(-1)^{\varepsilon_i}\operatorname{spr} X_i|$ is bounded by the spread of some observation $X_1, \ldots, X_n$. Consequently

$$\lim_{n\to\infty}\delta(\varepsilon_1,\ldots,\varepsilon_n) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(-1)^{\varepsilon_i}\operatorname{spr} X_i = 0,$$

so, by $\frac{1}{n}\sum_{i=1}^{n} x_i^{*} = \frac{1}{n}\sum_{i=1}^{n}\operatorname{mid} X_i + \frac{1}{n}\sum_{i=1}^{n}(-1)^{\varepsilon_i}\operatorname{spr} X_i$, the theorem is proved. □

Our method of finding endpoints of intervals that maximize $\overline{S^2}$ follows directly from Theorem 4.4. However, the result showed in this theorem is asymptotic, i.e. it is satisfying if the sample size is large enough. Therefore, in real life data analysis a natural question arises immediately: Is my sample large enough to believe that our method works satisfactorily? The answer depends, obviously, on the particular data set and is not universal even for samples of the same size. Below we specify the requirement a sample must satisfy so that the algorithm of identifying maximizers work correctly and return the true value of $\overline{S^2}$.

Keeping in mind some facts discussed in Section 4 we may conclude roughly that if a sample is large enough then the maximizing endpoints form two disjoint subsets separated by μ. Indeed, if n is large enough then m^*_{-i} is approximately equal to μ for any i. Conversely, if for at least one interval-valued observation X_i it happens that either $m^*_{-i} < \operatorname{mid} X_i < \mu$ or $\mu < \operatorname{mid} X_i < m^*_{-i}$ then the i-th maximizing endpoint could be determined incorrectly which implies that the underlying sample is not large enough. Therefore, it seems that our question about the number of observations required to assert that a sample is large enough can be replaced by a question on whether the actual distance between m^*_{-i} and μ is small enough to treat them as being equal. Thus let us try to assess $|m^*_{-i}-\mu|$. Obviously $|m^*_{-i}-\mu| \leqslant |m^*_{-i}-m^*|+|m^*-\mu|$. Moreover, by (10) we have

$$1+\frac{1}{n-1}=\frac{x^*_i-m^*_{-i}}{x^*_i-m^*}=1+\frac{m^*-m^*_{-i}}{x^*_i-m^*},$$

and hence $|m^*-m^*_{-i}|=\frac{1}{n-1}|x^*_i-m^*|$. We also have $|x^*_i-m^*| \leqslant \max\{|\mu-\underline{x}_i|,|\mu-\overline{x}_i|\}+|m^*-\mu|$. On the other hand (see the proof of Lemma 4.4) $|m^*-\mu|=|\delta(\varepsilon_1,\ldots,\varepsilon_n)|<\frac{1}{n}\operatorname{spr} X_k$ for a certain k. Thus, gathering all these remarks we obtain

$$|m^*_{-i}-\mu|<\frac{1}{n-1}\Big(\max\{|\mu-\underline{x}_i|,|\mu-\overline{x}_i|\}+\operatorname{spr} X_k\Big). \tag{12}$$

In other words, inequality (12) determines the condition which must be satisfied for any i so that the algorithm worked correctly. This very condition could be applied and interpreted in practice as follows: if for every $i=1,\ldots,n$ the center of the interval observation X_i does not belong to a certain region, i.e.

$$\operatorname{mid} X_i \notin \mu \pm \frac{1}{n-1}\Big(\max\{|\mu-\underline{x}_i|,|\mu-\overline{x}_i|\}+\operatorname{spr} X_k\Big), \tag{13}$$

then the suggested algorithm indicates which endpoint of each interval X_i should be taken to maximize the sample variance.

Please, notice, that actually the sample size does not need to be very numerous. Indeed, it is sufficient for the centers of the intervals to satisfy a certain condition and then we can pick the desired endpoints.

The only problem left unsolved to apply (13) is to determine $\operatorname{spr} X_k$. Obviously, we could choose the largest of the spreads, i.e. $\max\{\operatorname{spr} X_1,\ldots,\operatorname{spr} X_n\}$, which is surely safe but rather a conservative approach. On the other hand, if the smallest of the spreads could be taken, then the length of the interval given in (13) would be the smallest which leads to much more liberal requirement. However, it is not always feasible to select the smallest spread – one can easily show an appropriate counterexample. Thus we may try to discover an intermediate solution – neither too conservative nor too liberal – but always acceptable.

Like in point (iv) of the proof of Lemma 4.4 let us consider an interval $I=(0,\sum_{i=1}^n \operatorname{spr} X_i)$ which is a union of n intervals $I_1,\ldots,I_n$, such that the length of I_i is equal to the spread of observation X_i, i.e. $\|I_i\|=\operatorname{spr} X_i$ for $i=1,\ldots,n$. Recall also that $\operatorname{spr} X_k$ can be taken into (13) if and only if there exists a permutation π of $I_1,\ldots,I_n$ such that the center c of the interval I lies within $I_{\pi(k)}$. It is possible only if $\frac{\|I\|}{2}-\|I_{\pi(k)}\|<\sum_{i=1}^{k-1}\|I_{\pi(i)}\|<\frac{\|I\|}{2}$. For a given permutation π an interval $\bigcup_{i=1}^j I_{\pi(i)}$ will be called the *minimal coverage* of an interval J if and only if

$$\Big(0,\sum_{i=1}^{j}\|I_{\pi(i)}\|\Big)\cap J=J \quad\text{and}\quad \Big(0,\sum_{i=1}^{j-1}\|I_{\pi(i)}\|\Big)\cap J\subsetneq J.$$

Let us arrange intervals $I_1,\ldots,I_n$ with respect to their lengths. Such permutation of these intervals we will denote by $I_{1:n},\ldots,I_{n:n}$ where $\|I_{1:n}\|\leqslant\ldots\leqslant\|I_{n:n}\|$. We additionally assume that at least one inequality is strict (because if all intervals are of the same length there is no problem in determining the desired spread in (13)). Therefore, there exists the smallest m such that $\|I_{1:n}\|<\|I_{m:n}\|$. Hence let us consider all minimal coverages of the interval $\big(\frac{\|I\|}{2}-\|I_{m:n}\|,\frac{\|I\|}{2}\big)$ including $I_{1:n}$. In particular, there exists such minimal coverage of the form $I_M=I_{1:n}\cup\bigcup_{i=1}^{j-1}I_{\pi(i)}$. Now, if we remove $I_{1:n}$, then the following hold:

1. I_M was a minimal coverage, so $\sum_{i=1}^{j-1}\|I_{\pi(i)}\|<\frac{\|I\|}{2}$,
2. $\|I_{1:n}\|<\|I_{m:n}\|$, so $\sum_{i=1}^{j-1}\|I_{\pi(i)}\|>\frac{\|I\|}{2}-\|I_{m:n}\|$.

This proves that we may put the second smallest spread into (13) for $\operatorname{spr} X_k$ to obtain a valid region.

5 Algorithm for computing $\overline{S^2}$

Before we present our new algorithm for computing the upper bound of the sample variance let us summarize some basic facts that lie behind it.

- Given any interval $X_i = [\underline{x}_i, \overline{x}_i]$ its element that maximizes $\overline{S^2}$ is that endpoint $\underline{x}_i$ or $\overline{x}_i$ which is more distant from m^*_{-i}.
- The lower endpoints $x^*_i = \underline{x}_i$, $i = 1, \ldots, n$ that maximize the sample variance are all smaller than the upper endpoints $x^*_i = \overline{x}_i$.
- If n is large enough, then $m^*_{-i} \simeq m^* \simeq \mu = \frac{1}{n}\sum_{j=1}^{n} \operatorname{mid} X_j$ for any $i = 1, \ldots, n$.

Algorithm 5.1. *Given a sample of intervals $X_1 = [\underline{x}_1, \overline{x}_n], \ldots, X_n = [\underline{x}_n, \overline{x}_n] \in \mathcal{K}_c(\mathbb{R})$:*

1. *Find the centers of all intervals, i.e.* $\operatorname{mid} X_i = \frac{1}{2}(\underline{x}_i + \overline{x}_i)$, $i = 1, \ldots, n$.
2. *Compute the arithmetic mean of the interval centers, i.e.* $\mu = \frac{1}{n}\sum_{i=1}^{n} \operatorname{mid} X_i$.
3. *For $i = 1$ to n do: if* $\operatorname{mid} X_i < \mu$ *then* $x^*_i = \underline{x}_i$, *else* $x^*_i = \overline{x}_i$.
4. *Compute the upper bound of the sample variance* $\overline{S^2} = S^2(x^*_1, \ldots, x^*_n)$.

Remark 5.2. *One can see easily that the suggested algorithm works in $O(n)$ time.*

Below we present the `R` code to check whether the algorithm can be applied for a given data set and compute the maximum value of the variance (provided it can be done). An argument *data* is a data frame whose first column is a vector of centers and the second column is a vector of spreads.

```
maxVar = function(data){
n = dim(data)[1]
resp = numeric(n)
left = data[,1]-data[,2]
right = data[,1]+data[,2]
data1 = sort(data[,2])
secNum = which(diff(data1) != 0)[1]+1
sec = data1[secNum]
mid = mean(data[,1])
for(i in 1:n){
 reg1 = mid-(max(abs(mid-left[i]), abs(mid-right[i]))+sec)/(n-1)
 reg2 = mid+(max(abs(mid-left[i]), abs(mid-right[i]))+sec)/(n-1)
 resp[i] = ifelse(data[i,1] > reg1 & data[i,1] < reg2, 1, 0)
        }
if(sum(resp) == 0) {
 respAux = numeric(n)
 for(i in 1:n){ respAux[i] = sign(data[i,1]-mid)}
 obsAux = data[,1]+respAux*data[,2]
 cat("Maxiumum variance is equal to ", var(obsAux), "\n")
        }
else {cat("The algorithm cannot be applied to this data set.",
   "\n")}
}
```

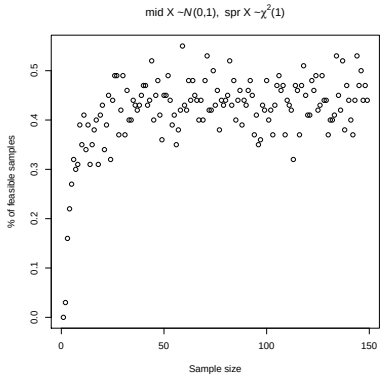
6 Simulation study

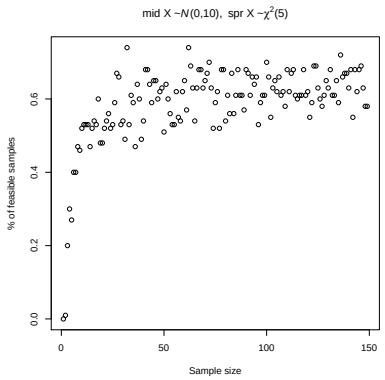
We illustrate some aspects of our algorithm by showing a few simulation results.
(i) Firstly, let us consider two small samples of size $n = 6$. In both cases we have drawn the centers of the intervals from the uniform distribution $\mathcal{U}(-10, 10)$ and the corresponding spreads from the exponential distribution $Exp(0.1)$. The generated intervals are shown in Table 1. One can check that the desired condition to apply the algorithm in case of sample 1 is fulfilled and we obtain $\overline{S^2} = 348.91$. However, the algorithm cannot be applied to sample 2. The last

Sample 1		Sample 2	
mid X	spr X	mid X	spr X
6.14	2.56	0.04	2.68
2.94	3.72	-7.78	6.85
1.66	0.99	-9.41	4.92
-8.93	4.96	2.09	1.35
-7.83	23.68	3.61	30.07
7.74	13.20	-4.95	4.39

Table 1: Small samples generated by: $\text{mid}\, X_i \sim \mathcal{U}(-10, 10)$ and $\text{spr}\, X_i \sim Exp(0.1)$.

conclusion is not surprising since the sample size is very small here. Anyway, sample 1 shows that it may happen that our algorithm works even for such small number of observations.
(ii) Now let us consider two bigger samples of size $n = 50$ with centers drawn from the normal distribution $N(-10, 10)$ and the corresponding spreads from the exponential distribution $Exp(1)$. The intervals are given in Table 2. The desired condition to apply the algorithm in case of sample 3 is satisfied and we obtain $\overline{S^2} = 40.11$. However, in the case of sample 4 the number of observations still is not large enough and the algorithm cannot be applied. Therefore, we see that the sample size usually considered as moderate sometimes may be sufficient to apply our algorithm but not necessarily. As we know, the algorithm proposed in Section 5 is the asymptotic one, so it works nicely for large samples. However, it is clear that the statement "large enough" is fuzzy and we would be satisfied if the algorithm could be applied wherever we need to estimate a variance of interval data. That is why in the next simulation we have tried to estimate how often it would be possible to apply our algorithm for different sample sizes. Therefore, we have considered interval samples of different sizes $n = 2, \ldots, 150$. For each fixed sample size n we have generated 100 samples and computed the fraction of samples for which the algorithm was feasible for application. We have conducted this experiment for various distributions, in particular: $\text{mid}\, X_i \sim N(0, 1)$ and $\text{spr}\, X_i \sim \chi^2(1)$, $\text{mid}\, X_i \sim N(0, 10)$ and $\text{spr}\, X_i \sim \chi^2(5)$, $\text{mid}\, X_i \sim N(0, 10)$ and $\text{spr}\, X_i \sim \chi^2(1)$. Some results are presented in the Figure 1. As it is seen the meaning of "large enough" sample size depends also on the sample distribution.





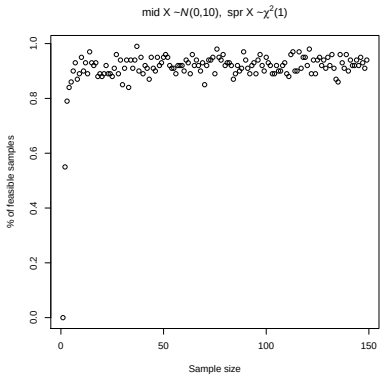


Figure 1: Fraction of samples for which the new algorithm works.

Sample 3				Sample 4			
mid X	spr X	mid X	spr X	mid X	spr X	mid X	spr X
2.32	0.20	8.03	1.63	1.04	1.13	-4.56	2.51
0.51	1.17	-0.26	1.65	-4.54	0.42	6.25	0.74
-3.79	0.96	-1.03	0.30	0.46	0.18	2.65	1.61
0.85	0.00	1.14	0.58	6.33	0.17	-0.60	0.26
2.87	0.05	-14.08	0.05	-4.35	2.06	-2.35	2.21
-4.36	0.69	-4.29	3.06	-2.74	0.99	-2.08	2.67
10.92	1.46	1.95	2.60	5.40	0.22	3.75	1.73
2.35	1.19	-0.45	0.65	2.58	2.05	0.62	3.89
-3.20	0.83	-8.13	0.45	4.45	0.08	-4.66	3.35
-11.82	1.63	-0.27	0.12	-5.20	0.42	2.80	1.17
-1.93	2.00	-7.03	0.24	-2.98	0.45	7.10	0.34
-9.12	1.20	2.53	0.43	-0.57	0.12	0.34	3.67
-5.72	1.31	-0.75	0.85	-1.76	1.21	0.12	3.19
-6.26	1.71	-2.96	0.21	-2.86	1.12	9.22	1.31
-4.48	0.81	-6.41	1.87	0.85	0.70	0.02	0.43
-3.42	1.25	10.00	0.42	-0.72	2.91	-0.49	0.44
2.67	0.09	3.11	1.36	-3.94	1.02	-8.84	0.16
-6.03	0.56	3.61	1.42	0.03	0.16	2.03	0.95
3.48	1.77	-3.60	0.14	6.31	0.49	5.52	0.36
-1.18	0.03	3.19	1.08	-3.90	1.10	1.68	0.66
4.86	0.88	0.51	1.92	5.70	0.81	6.04	0.66
-2.66	2.85	-11.36	1.05	-12.27	1.08	-0.43	3.78
-1.75	2.91	0.35	0.60	4.01	0.17	-1.98	1.13
-1.10	0.47	0.21	1.02	-0.95	1.36	8.96	1.16
-5.29	3.41	5.40	6.75	-1.71	1.11	-2.35	1.74

Table 2: Moderate samples generated by: $\operatorname{mid} X_i \sim \mathcal{N}(0, 10)$ and $\operatorname{spr} X_i \sim Exp(1)$.

(iii) In [26] the authors consider an example of real-life interval-valued data set *Car* of different models of cars produced worldwide and described by such variables as *Price, Engine capacity and Top speed*, among others (see Table 3).

For *Price* the new algorithm cannot be applied but for the other two it works well and produces the following values of the maximum variance: 3095342 and 2272.333 for *Engine capacity* and *Top speed*, respectively.

(iv) In the next simulation we generated 100 times samples of the size 1000 and checked that how many of them three different algorithms could be applied: our new algorithm, c_0-few intersection algorithm and no-subset property family algorithm. For c_0-few intersection a value $c_0 = 50$ was chosen which is 5% of the whole sample. We drew centers and spreads of intervals from different distribution which are listed in Table 4. First given is the distribution of centers and then the distribution of the spreads.

(v) The next experiment was performed to compare the applicability of our new algorithm with respect to those mentioned in Section 3, i.e. designed for samples satisfying the no-subset property or admitting fixed c_0-few intersections. A sample of the size n was drawn from $\operatorname{mid} X_i \sim N(0, 10)$ and $\operatorname{spr} X_i \sim \chi^2(1)$, where $n \in \{10, 150, 500\}$. For each n the experiment was repeated 100

	Price	Engine Capacity	Top Speed
Alfa 145	[27806, 33596]	[1370, 1910]	[185, 211]
Alfa 156	[41593, 62291]	[1598, 2492]	[200, 227]
Alfa 166	[64499, 88760]	[1970, 2959]	[204, 211]
Aston Martin	[260500, 460000]	[5935, 5935]	[298, 306]
Audi A3	[40230, 68838]	[1595, 1781]	[189, 238]
Audi A6	[68216, 140205]	[1781, 4172]	[216, 250]
Audi A8	[123849, 171417]	[2771, 4172]	[232, 250]
BMW 3	[45407, 76392]	[1796, 2979]	[201, 247]
BMW 5	[70292, 198792]	[2171, 4398]	[226, 250]
BMW 7	[104892, 276792]	[2793, 5397]	[228, 240]
Ferrari	[240292, 391692]	[3586, 5474]	[295, 298]
Punto	[19229, 30885]	[1242, 1910]	[155, 170]
Fiesta	[19242, 24742]	[1242, 1753]	[155, 170]
Skoda Fabia	[19519, 32686]	[1397, 1896]	[157, 183]
Skoda Octavia	[27419, 48679]	[1585, 1896]	[190, 191]
Passat	[39676, 63455]	[1595, 2496]	[192, 220]

Table 3: A subset of *Car* data set discussed in [26] with 16 car models and 3 variables.

	New algorithm	c_0 intersections	no subset
$\mathcal{N}(0,10), Exp(0.1)$	0.43	0	0
$\mathcal{N}(0,10), Exp(1)$	0.93	0	0
$\mathcal{U}(-1,1), Exp(1)$	0.43	0	0
$\mathcal{U}(-1,1), Exp(0.1)$	0	0	0
$\mathcal{U}(-1,1), \Gamma(1)$	0.36	0	0
$\mathcal{U}(-1,1), \Gamma(0.1)$	0.89	0	0
$\mathcal{N}(0,10), \Gamma(1)$	0.94	0	0
$\mathcal{N}(0,10), \Gamma(0.1)$	1	1	0
$\mathcal{P}(10), \mathcal{U}(0,10)$	0.94	0	0
$\mathcal{P}(10), Exp(0.1)$	0.65	0	0
$\mathcal{P}(10), Exp(1)$	0.99	1	0
$\mathcal{P}(10), \Gamma(1)$	0.97	0.99	0
$\mathcal{P}(10), \Gamma(0.1)$	1	1	0

Table 4: Fractions of samples for which algorithms worked.

times and a fraction of samples for which the algorithm was feasible for application was calculated for each algorithm. Results are given in Table 5. It shows undoubtedly that the new algorithm could be helpful in estimating the sample variance more often than its competitors.

(vi) Finally, in cases when all algorithms were feasible we measured the system time required for calculating $\overline{S^2}$. A few results for different sample sizes are shown in Table 6.

7 Conclusions

The sample variance computation under interval uncertainty is in general the NP-hard problem. That is why each efficient algorithm that leads to the solution in a feasible time needs some limitations on data. Usually these limitations are quite strong. Popular algorithms which work in polynomial time could be applied for a narrow class of intervals like those satisfying the no-subset property or c_0-few intersections only. This is the reason why an algorithm that could be effectively applied for a broader class of intervals is of interest. The new algorithm (based on asymptotic reasoning) introduced in this paper gives a concise and easily-verifiable conditions which must be satisfied by a data set so that the new algorithm could work correctly. It appears that the suggested algorithm which works in linear time could be a new competitive tool useful in statistical inference and data analysis based on interval-valued data helpful in various Big Data problems. Moreover, by the resolution identity, one can apply this algorithm for estimating the upper bound of a fuzzy sample variance. The problem considered in this contribution deserves further study. Firstly, it would be desirable to examine whether it is possible to improve condition (13) which guarantees that the suggested algorithm works correctly. Next direction of the future considerations is to modify this algorithm so it might be

Sample size	No subset property	c_0-few intersections	New algorithm
$n = 10$	0.09	0.60	0.89
$n = 150$	0	0.52	0.90
$n = 500$	0	0	0.92

Table 5: Fraction of feasible samples for different algorithms and different sample sizes.

Sample size	No subset property	c_0-few intersections	New algorithm
$n = 6$	< 0.00	< 0.00	< 0.00
$n = 150$	< 0.00	0.05	< 0.00
$n = 500$	0.01	0.16	0.01

Table 6: System times (in seconds) for different algorithms and different sample sizes.

applied for estimating correlation of two features described by interval-valued data.

References

[1] C. Angulo, D. Anguita, L. Gonzalez-Abril, J. A. Ortega, *Support vector machines for interval discriminant analysis*, Neurocomputing, **71** (2008), 1220-1229.

[2] J. Antoch, R. Miele, *Use of genetic algorithms when computing variance of interval data*, In: B. Fichet et al. (Eds.), Classification and Multivariate Analysis for Complex Data Structures, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, 2011.

[3] A. Blanco-Fernández, A. Colubi, M. García-Bárzana, *A set arithmetic-based linear regression model for modelling interval-valued responses through real-valued variables*, Information Sciiences, **247** (2013), 109-122.

[4] C. Cappelli, P. D'Urso, F. Di Iorio, *Regime change analysis of interval-valued time series with an application to PM10*, Chemometrics and Intelligent Laboratory Systems, **146** (2015), 337-346.

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

[6] E. Dantsin, V. Kreinovich, A. Wolpert, G. Xiang, *Population variance under interval uncertainty: a new algorithm*, Reliable Computing, **12** (2006), 273-280.

[7] P. D'Urso, L. De Giovanni, R. Massari, *Trimmed fuzzy clustering for interval-valued data*, Advances Data Analysis and Classification, **9** (2015), 21-40.

[8] P. D'Urso, P. Giordani, *A least squares approach to principal component analysis for interval valued data*, Chemometrics and Intelligent Laboratory Systems, **70** (2004), 179-192.

[9] A. P. Duarte Silva, P. Brito, *Discriminant analysis of interval data: An assessment of parametric and distance-based approaches*, Journal of Classification, **32** (2015), 516-541.

[10] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, M. Aviles, *Exact bounds on finite populations of interval data*, Reliable Computing, **11** (2005), 207-233.

[11] M. Gagolewski, *Spread measures and their relation to aggregation functions*, European Journal of Operational Research, **241** (2015), 469-477.

[12] A. Jalal-Kamali, V. Kreinovich, *Estimating correlation under interval uncertainty*, Mechanical Systems and Signal Processing, **37** (2013), 43-53.

[13] A. Kołacz, P. Grzegorzewski, *Measures of dispersion for multidimensional data*, European Journal of Operational Research, **251** (2016), 930-937.

[14] V. Kreinovich, S. Ferson, *Computing best-possible bounds for the distribution of a sum of several variables is NP-hard*, International Journal of Approximate Reasoning, **41** (2006), 331-342.

[15] V. Kreinovich, H. T. Nguyen, B. Wu, *On-line algorithms for computing mean and variance of interval data, and their use in intelligent systems*, Information Sciences, **177** (2007), 3228-3238.

[16] V. Kreinovich, G. Xiang, *Fast algorithms for computing statistics under interval uncertainty: An overview*, In: V. N. Huynh et al. (Eds.), Interval/Probabilistic Uncertainty and Non-Classical Logics, Springer, 2008, 19-31.

[17] V. Kreinovich, G. Xiang, S. Ferson, *Computing mean and variance under DempsterShafer uncertainty: Towards faster algorithms*, International Journal of Approximate Reasoning, **42** (2006), 212-227.

[18] R. E. Moore, *Interval Analysis*, Prentice-Hall, 1966.

[19] R. E. Moore, R. B. Kearfott, M. J. Cloud, *Introduction to Interval Analysis*, SIAM, 2009.

[20] H. T. Nguyen, V. Kreinovich, B. Wu, G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer, 2012.

[21] A. Oussous, F. Z. Benjelloun, A. A. Lahcen, S. Belfkih, *Big Data technologies: A survey*, Journal of King Saud University-Computer and Information Sciences, **30** (2018), 431-448.

[22] A. B. Ramos-Guajardo, A. Colubi, G. González-Rodríguez., *Inclusion degree tests for the Aumann expectation of a random interval*, Information Sciences, **288** (2014), 412-422.

[23] A. B. Ramos-Guajardo, P. Grzegorzewski, *Distance-based linear discriminant analysis for interval-valued data*, Information Sciences, **372** (2016), 591-60.

[24] B. Sinova, A. Colubi, M. A. Gil, G. González-Rodríguez, *Interval arithmetic-based linear regression between interval data: Discussion and sensitivity analysis on the choice of the metric*, Informtion Sciences, **199** (2012), 109-124.

[25] A. Skowron, A. Jankowski, S. Dutta, *Interactive granular computing*, Granular Computing, **1** (2016), 95-113.

[26] R. M. C. R. Souza, D. C. F. Queiroz, F. J. A. Cysneiros, *Logistic regression-based pattern classifiers for symbolic interval data*, Pattern Analysis and Applications, **14** (2011), 273-282.

[27] T. Sunaga, *Theory of interval algebra and its application to numerical analysis*, RAAG Memoirs, Ggujutsu Bunken Fukuykai, Tokyo, **2** (1958), 29-46, 547-564.

[28] S. A. Vavasis, *Nonlinear Optimization: Complexity Issues*, Oxford University Press, New York, 1991.

[29] H. Wang, Z. Xu, H. Fujita, S. Liu, *Towards felicitous decision making: An overview on challenges and trends of Big Data*, Information Sciences, **367-368** (2016), 747-765.

[30] M. Warmus, *Calculus of approximations*, Bulletin de l'Academie Polonaise de Sciences, **4** (1956), 253-257.

[31] G. Xiang, M. Ceberio, V. Kreinovich, *Computing population variance and entropy under interval uncertainty: linear-time algorithms*, Reliable Computing, **13** (2007), 467-488.