On Statistics, Probability, and Entropy of Interval-Valued Datasets

Chenyi Hu^{1(✉)} and Zhihui H. Hu²

¹ University of Central Arkansas, Conway, AR, USA
chu@uca.edu

² Edge Pursuit LLC, San Francisco, CA, USA

Abstract. Applying interval-valued data and methods, researchers have made solid accomplishments in information processing and uncertainty management. Although interval-valued statistics and probability are available for interval-valued data, current inferential decision making schemes rely on point-valued statistic and probabilistic measures mostly. To enable direct applications of these point-valued schemes on interval-valued datasets, we present point-valued variational statistics, probability, and entropy for interval-valued datasets. Related algorithms are reported with illustrative examples.

Keywords: Interval-valued dataset · Point-valued variational statistics · Probability · Information entropy

1 Introduction

1.1 Why Do We Study Interval-Valued Datasets?

Statistic and probabilistic measures play a very important role in processing data and managing uncertainty. In the literature, these measures are mostly point-valued and applied to point-valued dataset. While a point-valued datum intends to record a snapshot of an event instantaneously in theory, it is often imprecise in real world due to system and random errors. Applying interval-valued data to encapsulate variations and uncertainty, researchers have developed interval methods for knowledge processing. With data aggregation strategies [1, 5, 21], and others, we are able to reduce large size point-valued data into smaller interval-valued ones for efficient data management and processing. By doing so, researchers are able to focus more on qualitative properties and ignore insignificant quantitative differences.

Studying interval-valued data, Gioia and Lauro developed interval-valued statistics [4] in 2005. Lodwick and Jamison discussed interval-valued probability [17] in the analysis of problems containing a mixture of possibilistic, probabilistic, and interval uncertainty in 2008. Billard and Diday reported regression analysis of interval-valued data in [2]. Huynh *et al.* established a justification on decision making under interval uncertainty [13]. Works on applications of interval-valued data in knowledge processing include [3, 8, 16, 19, 20, 22], and

many more. Applying interval-valued data in the stock market forecasting, Hu and He initially reported an astonishing quality improvements in [9]. Specifically, comparing against the commonly used point-valued confidence interval predictions, the interval approaches have increased the average accuracy ratio of annual stock market forecasts from 12.6% to 64.19%, and reduced the absolute mean error from 72.35% to 5.17% [9]. Additional results on the stock market forecasts reported in [6, 7, 10], and others have verified the advantages of using interval-valued data. The paper [12], published in the same volume as this one, further validates the advantages from the perspective of information theory.

Using interval-valued data can significantly improve efficiency and effectiveness in information processing and uncertainty management. Therefore, we need to study interval-valued datasets.

1.2 The Objective of this Study

As a matter of fact, powerful inferential decision making schemes in the current literature use point-valued statistic and probabilistic measures, not interval-valued ones [4] and [17], mostly. To enable direct applications of these schemes and theory on analyzing interval-valued datasets, we need to supply *point-valued* statistics and probability for *interval-valued* datasets. Therefore, the primary objective of this work is to establish and to calculate such point-valued measures for interval-valued datasets.

To make this paper easy to read, it includes brief introductions on necessary background information. It also provides easy to follow illustrative examples for novel concepts and algorithms in addition to pseudo-code. Numerical results of these examples are obtained with a recent version of Python 3. However, readers may use any preferred general purpose programming language to verify the results.

1.3 Basic Concepts and Notations

Prior to our discussion, let us first clarify some basic concepts and notations related to intervals in this paper. An *interval* is a connected subset of $\mathbb{R}$. We denote an interval-valued object with a boldfaced letter to distinguish it from a point-valued one. We further specify the greatest lower bound and least upper bound of an interval object with an underline and an overline of the same letter but not boldfaced, respectively. For example, while a is a real, the boldfaced letter $\mathbf{a}$ denotes an interval with its greatest lower bound $\underline{a}$, and least upper bound $\bar{a}$. That is $\mathbf{a} = \{a : \underline{a} \leq a \leq \bar{a}, a \in \mathbb{R}\} = [\underline{a}, \bar{a}]$. The absolute value of $\mathbf{a}$, defined as $|\mathbf{a}| = \bar{a} - \underline{a}$, is also called the length (or norm) of $\mathbf{a}$. This is the greatest distance between any two numbers in $\mathbf{a}$.

The *midpoint* and *radius* of an interval $\mathbf{a}$ are defined as $mid(\mathbf{a}) = \frac{\underline{a} + \bar{a}}{2}$ and $rad(\mathbf{a}) = \frac{\bar{a} - \underline{a}}{2}$, respectively. Because the midpoint and radius of an interval $\mathbf{a}$ are point-valued, we simply denote them as $mid(a)$ and $rad(a)$ without boldfacing the letter a . We call $[\underline{a}, \bar{a}]$ the endpoint (or min-max) representation of

a. We can specify an interval $\mathbf{a}$ with $\text{mid}(\mathbf{a})$ and $\text{rad}(\mathbf{a})$ too. This is because of $\underline{a} = \text{mid}(\mathbf{a}) - \text{rad}(\mathbf{a})$ and $\bar{a} = \text{mid}(\mathbf{a}) + \text{rad}(\mathbf{a})$. In the rest of this paper, we use both min-max and mid-rad representations for an interval-valued object.

While we use a boldfaced lowercase letter to indicate an interval, we denote an *interval-valued dataset*, i.e., a collection of real intervals, with a boldfaced uppercase letter. For instance, $\mathbf{X} = \{x_1, x_2, \dots, x_n\}$ is an interval-valued dataset. The sets $\underline{X} = \{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n\}$ and $\bar{X} = \{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n\}$ are the left- and right-end sets of $\mathbf{X}$, respectively. Although items in a set are not ordered, the $\underline{x}_i \in \underline{X}$ and $\bar{x}_i \in \bar{X}$ are related to the same interval $x_i \in \mathbf{X}$. For convenience, we denote both $\underline{X}$ and $\bar{X}$ as ordered tuples. They are the *left- and right-endpoints* of $\mathbf{X}$. That is $\underline{X} = (\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n)$ and $\bar{X} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)$. Similarly, the *midpoint and radius* of $\mathbf{X}$ are point-valued tuples. They are $\text{mid}(\mathbf{X}) = (\text{mid}(x_1), \text{mid}(x_2), \dots, \text{mid}(x_n))$ and $\text{rad}(\mathbf{X}) = (\text{rad}(x_1), \text{rad}(x_2), \dots, \text{rad}(x_n))$, respectively.

Example 1. Provided an interval-valued sample dataset $\mathbf{X}_0 = \{[1, 5], [1.5, 3.5], [2, 3], [2.5, 7], [4, 6]\}$. Then, its left-endpoint is $\underline{X}_0 = (1, 1.5, 2, 3, 2.5, 4)$, and right-endpoint is $\bar{X}_0 = (5, 3.5, 3, 7, 6)$. The midpoint of $\mathbf{X}_0$ is $\text{mid}(\mathbf{X}_0) = \frac{\underline{X}_0 + \bar{X}_0}{2} = (3, 2.5, 2.5, 4.75, 5)$, and the radius is $\text{rad}(\mathbf{X}_0) = \frac{\bar{X}_0 - \underline{X}_0}{2} = (2, 1, 0.5, 2.25, 1)$.

We use this sample dataset $\mathbf{X}_0$ in the rest of this paper to illustrate concepts and algorithms for its simplicity.

In the rest of this paper, we discuss statistics of an interval-valued dataset in Sect. 2; define point-valued probability distributions for an interval-valued dataset in Sect. 3; introduce point-valued information entropy in Sect. 4; and summarize the main results and future work in Sect. 5.

2 Descriptive Statistics of an Interval-Valued Dataset

We introduce positional statistics for an interval-valued dataset first, and then discuss its point-valued variance and standard deviation.

2.1 Positional Statistics of an Interval-Valued Dataset $\mathbf{X}$

The left-, right-endpoints, midpoint, and radius $\underline{X}, \bar{X}, \text{mid}(\mathbf{X})$, and $\text{rad}(\mathbf{X})$ are among positional statistics of an interval-valued dataset $\mathbf{X}$ as presented in Example 1. The mean of $\mathbf{X}$, denoted as $\mu_{\mathbf{X}}$, is the arithmetic average of $\mathbf{X}$. Because $\sum_{i=1}^n x_i = [\sum_{i=1}^n \underline{x}_i, \sum_{i=1}^n \bar{x}_i]$ in interval arithmetic¹, we have

$$\mu_{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n x_i = \left[\frac{\sum_{i=1}^n \underline{x}_i}{n}, \frac{\sum_{i=1}^n \bar{x}_i}{n} \right] = [\mu_{\underline{X}}, \mu_{\bar{X}}] \quad (1)$$

We now define few more observational statistics for $\mathbf{X}$.

¹ For readers who want to know more about standardized interval arithmetic, please refer the IEEE Standards for Interval Arithmetic [14] and [15].

Definition 1. Let $\mathbf{X}$ be an interval-valued dataset, then

1. The envelope of $\mathbf{X}$ is the interval $env(\mathbf{X}) = [\min(\underline{X}), \max(\overline{X})]$;
2. The core of $\mathbf{X}$ is the interval $core(\mathbf{X}) = \bigcap_{i=1}^n x_i = [\max(\underline{X}), \min(\overline{X})]$; and
3. The mode of $\mathbf{X}$ is a tuple, $mode(\mathbf{X}) = (\bigcap_{s \in S_j} x_s, k)$, where $\bigcap_{s \in S_j} x_s \neq \emptyset$, S_j is a cardinality k subset of $\{1, 2, \dots, n\}$, and for any $S_i \subseteq \{1, 2, \dots, n\}$ if $\bigcap_{s \in S_i} x_s \neq \emptyset$ then $|S_i| \leq k$.

In other words, $\forall x_i \in \mathbf{X}$, x_i is a subset of $env(\mathbf{X})$, and $core(\mathbf{X})$ is a subset of x_i . Furthermore, $mode(\mathbf{X})$ is an ordered tuple. In which, $\bigcap_{s \in S_j} x_s$ is the non-empty intersection of x_s for all $s \in S_j$, such that, the cardinality of S_j is the greatest. For a given $\mathbf{X}$, its mode may not be unique. This is because of that, there may be multiple cardinality k subsets of $\{1, 2, \dots, n\}$ satisfying the nonempty intersection requirement $\bigcap_{s \in S_j} x_s \neq \emptyset$.

Corollary 1. Let $\mathbf{X}$ be an interval-valued dataset, then

1. For all $x_i \in \mathbf{X}$, $x_i \subseteq env(\mathbf{X})$;
2. The core of $\mathbf{X}$ is not empty if and only if $\max(\underline{X}) \leq \min(\overline{X})$; and
3. The mode of $\mathbf{X}$ is $(core(\mathbf{X}), n)$ if and only if $core(\mathbf{X}) \neq \emptyset$.

Corollary 1 is straightforward.

Instead of providing a proof, we provide the mean, envelop, core and mode for the sample dataset $\mathbf{X}_0 = \{[1, 5], [1.5, 3.5], [2, 3], [2.5, 7], [4, 6]\}$. In addition to its endpoints, midpoint, and radius presented in Example 1, we have its mean $\mu_{\mathbf{X}_0} = [2.2, 4.9]$; $env(\mathbf{X}_0) = [1, 7]$; $core(\mathbf{X}_0) = \emptyset$ because of $\max(\underline{X}_0) = 4$ is greater than $\min(\overline{X}_0) = 3$; and $mode(\mathbf{X}_0) = ([2.5, 3], 4)$. Figure 1 illustrates the sample dataset $\mathbf{X}_0$. From which, one may visualize the $env(\mathbf{X}_0)$ and $mode(\mathbf{X}_0)$ by imaging a vertical line, like the y-axis, continuously moving from left to right. The first and last points the line touches any $x_i \in \mathbf{X}_0$ determine the envelop $env(\mathbf{X}_0) = [1, 7]$. The line touches at most four intervals for all $x_i \in \mathbf{X}_0$ between $[2.5, 3]$. Hence, the mode is $mode(\mathbf{X}_0) = ([2.5, 3], 4)$.

While finding the envelop, core, and mean of $\mathbf{X}$ is straightforward, determining the mode of $\mathbf{X}$ involves the $2n$ numbers in $\underline{X}$ and $\overline{X}$, which divide $env(\mathbf{X})$ into $2n - 1$ sub-intervals in general (though some of them maybe degenerated as points.) Each of these $2n - 1$ sub-intervals can be a candidate of the nonempty intersection part in the mode. For any $x_i \in \mathbf{X}$, it may cover some of these $2n - 1$ sub-intervals (candidates) consecutively. For each of these candidates, we accumulate its occurrences in each $x_i \in \mathbf{X}$. The mode(s) for $\mathbf{X}$ is (are) the candidate(s) with the (same) highest occurrence. As a special case, if $core(\mathbf{X})$ is not empty, then $mode(\mathbf{X}) = (core(\mathbf{X}), n)$. We summarize the above as an algorithm.

Algorithm 1: (Finding the mode for an interval dataset $\mathbf{X}$)

Input: $\mathbf{X}$: an n -element interval dataset.

Output: $mode(\mathbf{X})$.

If $\max(\underline{X}) < \min(\overline{X})$

```

mode(X) = ([max(X), min(X)], n)
Else
# Initialization:
Concatenating X and  $\overline{X}$  as a single list c
Sort the list c.
For i from 0 to  $2n - 1$ :
     $cand_i = ([c_i, c_{i+1}], count_i = 0)$ 
End for
# Counting frequency:
For each  $x_i \in \mathbf{X}$ 
    Find j and k, such that  $c_j = \underline{x}_i$  and  $c_k = \overline{x}_i$ 
    For l from j to k:
         $cand_l.count_l += 1$ 
    End for
End for
# Find the mode:
 $m = \max\{cand.count\}$ 
For j from 0 to  $2n - 1$ :
    If  $cand_j.count_j = m$ ,
         $mode(\mathbf{X}) = ([c_j, c_{j+1}], m)$ 
End for
Return  $mode(\mathbf{X})$ .

```

Algorithm 1 is $O(n^2)$. This is because of that for each interval x_i , it may update the count in each of the $2n - 1$ candidates takes $O(n^2)$.

Fig. 1. The sample interval-valued dataset $\mathbf{X}_0$.

2.2 Point-Valued Variational Statistics of an Interval-Valued Dataset

In the literature, the variance of a point-valued dataset X is defined as

$$Var(X) = \frac{1}{n} \sum_{i=1}^n |x_i - \mu|^2 \quad (2)$$

in which, the term $|x_i - \mu|$ is the distance between $x_i \in X$ and μ , which is the mean of X .

Using (2) to define a variance for an interval-valued $\mathbf{X}$, we need a notion of point-valued distance between two intervals, $x_i \in \mathbf{X}$ and the interval $\mu_{\mathbf{X}}$. May we simply use $|\mathbf{a} - \mathbf{b}|$, the absolute value of the difference between two intervals $\mathbf{a}$ and $\mathbf{b}$, as their distance? Unfortunately, it does not work.

In interval arithmetic [18], the difference between two intervals $\mathbf{a}$ and $\mathbf{b}$ is defined as the follow:

$$\mathbf{a} - \mathbf{b} = [\min\{\underline{a} - \underline{b}, \underline{a} - \bar{b}, \bar{a} - \underline{b}, \bar{a} - \bar{b}\}, \max\{\underline{a} - \underline{b}, \underline{a} - \bar{b}, \bar{a} - \underline{b}, \bar{a} - \bar{b}\}] \quad (3)$$

Equation (3) ensures $\forall a \in \mathbf{a}, \forall b \in \mathbf{b}, a - b \in \mathbf{a} - \mathbf{b}$. However, it also implies $|\mathbf{a} - \mathbf{b}| = \max\{|a - b|, \forall a \in \mathbf{a}, \forall b \in \mathbf{b}\}$, which is the maximum distance between $a \in \mathbf{a}$ and $b \in \mathbf{b}$.

Mathematically, a distance between two nonempty sets A and B is usually defined as the minimum distance between $a \in A$ and $b \in B$ but not the maximum. Hence, we need to define a notion of distance between two intervals.

Definition 2. Let $\mathbf{a}$ and $\mathbf{b}$ be two nonempty intervals. The distance between $\mathbf{a}$ and $\mathbf{b}$ is defined as

$$dist(a, b) = |mid(a) - mid(b)| + |rad(a) - rad(b)| \quad (4)$$

Definition 2 satisfies all mathematical requirements for a distance. They are $dist(a, b) \geq 0$; $dist(a, b) = 0$ if and only if $\mathbf{a} = \mathbf{b}$; $dist(a, b) = dist(b, a)$; and for any nonempty intervals $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$, $dist(a, c) \leq dist(a, b) + dist(b, c)$. Definition 2 is in fact an extension of the distance between two reals. This is because of that the radius of a real is zero and the midpoint of a real is itself always.

Replacing $x_i - \mu$ in Equation (2) with $dist(x_i, \mu_{\mathbf{X}})$ defined in (4), we have the point-valued variance of $\mathbf{X}$ as the follow:

$$\begin{aligned} Var(X) &= \frac{1}{n} \sum_{i=1}^n dist^2(x_i, \mu_{\mathbf{X}}) = \sum_{i=1}^n [|mid(x_i) - mid(\mu_{\mathbf{X}})| + |rad(x_i) - rad(\mu_{\mathbf{X}})|]^2 \\ &= \frac{1}{n} \sum_{i=1}^n (|mid(x_i) - mid(\mu_{\mathbf{X}})|)^2 + \frac{1}{n} \sum_{i=1}^n (|rad(x_i) - rad(\mu_{\mathbf{X}})|)^2 \\ &\quad + \frac{2}{n} \sum_{i=1}^n (|mid(x_i) - mid(\mu_{\mathbf{X}})|)(|rad(x_i) - rad(\mu_{\mathbf{X}})|). \end{aligned}$$

The expression above has three terms. All of them involve $mid(\mu_{\mathbf{X}})$ and $rad(\mu_{\mathbf{X}})$.

Since $\mu_{\mathbf{X}} = \left[\frac{\sum_{i=1}^n \underline{x}_i}{n}, \frac{\sum_{i=1}^n \bar{x}_i}{n} \right]$, $mid(\mu_{\mathbf{X}}) = \frac{1}{2} \left(\sum_{i=1}^n \underline{x}_i/n + \sum_{i=1}^n \bar{x}_i/n \right)$
 $= \frac{1}{n} \sum_{i=1}^n \left(\frac{\underline{x}_i + \bar{x}_i}{2} \right) = \frac{1}{n} \sum_{i=1}^n mid(x_i) = \mu_{mid(X)}$. Therefore, the first term in the
 expression above $\frac{1}{n} \sum_{i=1}^n (|mid(x_i) - mid(\mu_{\mathbf{X}})|)^2 = \frac{1}{n} \sum_{i=1}^n (mid(x_i) - \mu_{mid(X)})^2 =$
 $Var(mid(X))$ according to (2). Similarly, the second term $\frac{1}{n} \sum_{i=1}^n (|rad(x_i) -$
 $rad(\mu_{\mathbf{X}})|)^2 = Var(rad(X))$.

The third term is related to the absolute covariance between $mid(X)$ and $rad(X)$. Let $\Delta m_i = mid(x_i) - mid(\mu_{\mathbf{X}})$ and $\Delta r_i = rad(x_i) - rad(\mu_{\mathbf{X}})$, then
 we can rewrite the term $\frac{2}{n} \sum_{i=1}^n (|mid(x_i) - \mu_{mid(X)}|)(|rad(x_i) - \mu_{rad(X)}|)$ as
 $\frac{2}{n} \sum_{i=1}^n |\Delta m_i \Delta r_i|$.

Summarizing the discussion above, we have the point-valued variance for an interval-valued dataset $\mathbf{X}$ as the follow.

Definition 3. Let $\mathbf{X} = (x_1, x_2, \dots, x_n)$ be an interval-valued dataset, then the point-valued variance of $\mathbf{X}$ is

$$Var(X) = Var(mid(X)) + Var(rad(X)) + \frac{2}{n} \sum_{i=1}^n |\Delta m_i \Delta r_i| \quad (5)$$

Because midpoints and radii of interval-valued objects are point-valued, the variance defined in (5) is also point-valued. Hence, we have the point-valued standard deviation of $\mathbf{X}$ as usual:

$$Std(X) = \sqrt{Var(X)} \quad (6)$$

In evaluating (5) and (6), one does not need interval computing at all. For the sample dataset $\mathbf{X}_0$, we have its point-valued variance $Var(X_0) =$
 $var(mid(X_0)) + var(rad(X_0)) + \frac{2}{5} \sum_{i=1}^5 |\Delta m_i \Delta r_i| = 1.5125 + 0.55 + 1.282 = 3.3445;$
 and the standard deviation $Std(X_0) = 1.8288$.

It is worthwhile to note that, Eq. (5) is an extension of (2) and applicable to point-valued datasets too. This is because of that, for all x_i in a point-valued X , $rad(x_i) = 0$ and $mid(x_i) = x_i$ always. Hence, $Var(X) = Var(mid(X))$ for a point-valued X .

3 Probability Distributions of an Interval-Valued Population

An interval-valued dataset $\mathbf{X}$ can be viewed as a sample of an interval-valued population. In this section, we study practical ways to find probability distributions for an interval-valued dataset $\mathbf{X}$. Our discussion addresses two different cases. One assumes distribution information for all $x_i \in \mathbf{X}$. The other does not.

3.1 On Probability Distribution of $\mathbf{X}$ with Distribution Information for Each $x_i \in \mathbf{X}$

Our discussion involves the concept of a probability distribution over an interval. Let us very briefly review the literature first.

A function $f(x)$ is a probability density function (*pdf*) of a random variable x on the interval $x = [\underline{x}, \bar{x}]$ if and only if $f(x) \geq 0, \forall x \in \mathbb{R}$, and $\int_{-\infty}^{\infty} f(t)dt =$

$\int_{\underline{x}}^{\bar{x}} f(t)dt = 1$. Well-known *pdfs* in the literature include the uniform distribution:

$f(x) = \begin{cases} 1 & \text{if } x \in [0, 1], \\ 0 & \text{otherwise.} \end{cases}$; normal distribution: $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$;

and beta distribution: $f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$, where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and

both parameters α and β are positive, and $\Gamma(t)$ is the gamma function. There are software tools available to fit point-valued sample data, which means computationally determining the parameter values in a chosen type of distribution. For instance, the Python `scipy.stats` module is available to find the optimal μ and σ to fit a point-valued dataset in a normal distribution, and/or α and β in a beta distribution.

It is safe to assume an availability of a *pdf* for each $x_i \in \mathbf{X}$ both theoretically and computationally. In practice, an interval $x_i \in \mathbf{X}$ is often obtained through aggregating observed points. For instances, in [9] and [11], min-max and confidence intervals are applied to aggregate points into intervals, respectively. If an interval is provided directly, one can always pick points from the interval and fit these points with a selected probability distribution computationally. Hereafter, we denote the *pdf* of $x_i \in \mathbf{X}$ as $pdf_i(x)$.

We now define a notion of *pdf* for an interval-valued dataset $\mathbf{X}$.

Definition 4. A function $f(x)$ is called a probability density function of an interval-valued dataset $\mathbf{X} = \{x_1, x_2, \dots, x_n\}$ if and only if $f(x)$ satisfies all of the conditions:

$$\begin{cases} f(x) \geq 0 \ \forall x \in (-\infty, \infty); \\ \sum_{i=1}^n \int_{x_i \in \mathbf{X}} f(t)dt = 1. \end{cases} \quad (7)$$

The theorem below provides a practical way to calculate a *pdf* for $\mathbf{X}$.

Theorem 1. Let $\mathbf{X} = (x_1, x_2, \dots, x_n)$ be an interval-valued dataset; and $pdf_i(x)$ be the pdf of x_i provided $i \in \{1, 2, \dots, n\}$. Then,

$$f(x) = \frac{\sum_{i=1}^n pdf_i(x)}{n} \quad (8)$$

is a pdf of $\mathbf{X}$.

Proof. Because $pdf_i(x) \geq 0 \forall i \in \{1, 2, \dots, n\}$, we have $\sum_{i=1}^n pdf_i(x) \geq 0$. Hence,

$f(x) \geq 0$. In addition, $\int_{-\infty}^{\infty} pdf_i(t)dt = 1$ for all $i \in \{1, 2, \dots, n\}$, we have

$$\sum_{i=1}^n \int_{x_i} f(t)dt = \int_{-\infty}^{\infty} \frac{\sum_{i=1}^n pdf_i(x)}{n} dx = \frac{\sum_{i=1}^n \int_{-\infty}^{\infty} pdf_i(t)dt}{n} = \frac{n}{n} = 1.$$
 Equation (7) satisfied. Hence, the $f(x)$ is a pdf of $\mathbf{X}$. $\square$

Equation (8) actually provides a practical way of calculating the pdf of $\mathbf{X}$. Provided $pdf_i(x)$ for each $x_i \in \mathbf{X}$, we have the algorithm in pseudo-code below:

Algorithm 2: (Finding a pdf for $\mathbf{X}$)

Input: an n -item interval-valued dataset $\mathbf{X}$;

$pdf_i(x)$ for every $x_i \in \mathbf{X}$

Output: $pdf(\mathbf{X})$

Initialization:

Concatenating $\underline{x}$ and $\bar{x}$ as a list c

Sort c

For i from 1 to $2n - 1$:

$segment_i = (c_i, c_{i+1}, 0)$

End for

Accumulating pdf on each segment:

For each $x_i \in \mathbf{X}$ find the j and k , such that

$c_j = \underline{x}_i$ and $c_k = \bar{x}_i$

For l from j to k :

$segment_l.pdf += pdf_i$

End for

End for

Calculating the pdf:

For i from 0 to $2n - 1$:

$segment_i.pdf /= n$

End for

Return $segment_i$ for all $i \in \{1, 2, \dots, 2n - 1\}$

Example 2. Find a pdf from the sample dataset $\mathbf{X}_0 = \{[1, 5], [1.5, 3.5], [2, 3], [2.5, 7], [4, 6]\}$. For simplicity, we assume a uniform distribution for each pdf_i 's, i.e.,

$$pdf_i(x) = \begin{cases} \frac{1}{\bar{x}_i - \underline{x}_i} & \text{if } x \in x_i \\ 0, & \text{otherwise.} \end{cases}$$

Applying Algorithm 2, we have

$$f(X_0) = \frac{\sum_{i=1}^5 pdf_i(x)}{5} = \begin{cases} 0.05 & \text{if } x \in [1, 1.5] \\ 0.15 & \text{if } x \in (1.5, 2] \\ 0.35 & \text{if } x \in (2, 2.5] \\ 0.39 & \text{if } x \in (2.5, 3] \\ 0.19 & \text{if } x \in (3, 3.5] \\ 0.09 & \text{if } x \in (3.5, 4] \\ 0.19 & \text{if } x \in (4, 5] \\ 0.14 & \text{if } x \in (5, 6] \\ 0.044 & \text{if } x \in (6, 7] \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

The *pdf* in the example is a stair function. This is because the uniform distribution assumption on each $x_i \in \mathbf{X}$. $\square$

Here are few additional notes on finding a *pdf* for $\mathbf{X}$ with Algorithm 2 .

If assuming uniform distribution, how do we handle the case if $\exists i$ such that $\underline{x}_i = \bar{x}_i$? First of all, an interval element $\mathbf{x}_i$ is usually not degenerated as a constant. Even there is an i such that $\underline{x}_i = \bar{x}_i$, we can always assign an arbitrary non-negative *pdf* value at that point. This does not impact the calculation of probability in integrating the *pdf* function.

Algorithm 2 assumes $pdf_i(x) = 0, \forall x \notin x_i$. If it is not the case, the $2n$ numbers in $\underline{X}$ and $\bar{X}$ divide $\mathbb{R}$ in $2n + 1$ sub-intervals. They are $(-\infty, \min(\underline{X}))$, $(\max(\bar{X}), \infty)$ together with the $2n - 1$ sub-intervals in $env(\mathbf{x})$. Therefore, the accumulation loop in Algorithm 2 should run through all of the $2n + 1$ sub-intervals, and then normalize them by dividing n .

Another implicit assumption of Theorem 1 is that, all $x_i \in \mathbf{X}$ are equally weighted. However, that is not necessary. If needed, one may place a positive weight w_i on each of pdf_i 's as stated in the Corollary 2.

Corollary 2. *Let $\mathbf{X} = (x_1, x_2, \dots, x_n)$ be an interval-valued dataset and pdf_i be the *pdf* of $x_i \in X$, then the function*

$$f(x) = \frac{\sum_{i=1}^n w_i pdf_i(x)}{\sum_{i=1}^n w_i} \quad \text{where } \forall i w_i > 0 \quad (10)$$

is a pdf of $\mathbf{X}$.

A proof of Corollary 2 is straightforward too. We have successfully applied the Corollary in computationally studying the stock market [12].

3.2 Probability Distribution of an Interval-Valued $\mathbf{X}$ Without Distribution Information for Any $x_i \in \mathbf{X}$

It is not necessary to assume the probability distribution for all $x_i \in \mathbf{X}$ to find a *pdf* of $\mathbf{X}$. An interval x is determined by its midpoint and radius. Let $u = mid(x)$ and $v = rad(x)$ be two point-valued random variables. Then, the *pdf* of x is a

non-negative function $f(u, v) \geq 0$, such that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(u, v) du dv = 1$. If we assume a normal distribution for $au + bv$, then $f(u, v)$ is a bivariate normal distribution [25]. The *pdf* of a bivariate normal distribution is:

$$p(u, v) = \frac{1}{2\pi\sigma_u\sigma_v\sqrt{1-\rho^2}} e^{\frac{-z}{2(1-\rho^2)}} \quad (11)$$

where $z = \frac{(u - \mu_u)^2}{\sigma_u^2} - \frac{2\rho(u - \mu_u)(v - \mu_v)}{\sigma_u\sigma_v} + \frac{(v - \mu_v)^2}{\sigma_v^2}$ and ρ is the normalized correlation between u and v , i.e., the ratio of their covariance and the product of σ_u and σ_v . Applying the *pdf*, we are able to estimate the probability over a region $u = [u_1, u_2]$, $v = [v_1, v_2]$ as

$$P(x) = \int_{v_1}^{v_2} \int_{u_1}^{u_2} p(u, v) du dv \quad (12)$$

To calculate the probability of an interval x , whose midpoint and radius are u_0 and v_0 , we need a marginal *pdf* for either u or v . If we fix $u = u_0$, then the marginal *pdf* of v follows a single variable normal distribution. Thus,

$$p(v) = \frac{1}{\sigma_v\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{v - \mu_v}{\sigma_v})^2}, \quad (13)$$

and the probability of x is

$$P(x) = \int_{-v_0}^{v_0} p(v) dv \quad (14)$$

An interval-valued dataset $\mathbf{X}$ provides us its $mid(X)$ and $rad(X)$. They are point-valued sample sets of u and v , respectively. All of $\mu_{mid(X)}$, $\mu_{rad(X)}$, $\sigma_{mid(X)}$, and $\sigma_{rad(X)}$ can be calculated as usual to estimate the μ_u , μ_v , σ_u , and σ_v in (11). For instance, from the sample $\mathbf{X}_0$, we have $\mu_{mid(X_0)} = 3.55$, $\mu_{rad(X_0)} = 1.35$, $\sigma_{mid(X_0)} = 1.1$, $\sigma_{rad(X_0)} = 0.66$, and $\rho = 0.404$, respectively. Furthermore, using $\mu_{rad(X_0)} = 1.35$ and $\sigma_{rad(X_0)} = 0.66$ in (13), we can estimate the probability of an arbitrary interval x with (14).

So far, we have established practical ways to calculate point-valued variance, standard deviation, and probability distribution for an interval-valued dataset X . With them, we are able to *directly apply commonly available inferential decision making schemes based on interval-valued dataset*.

4 Information Entropy of Interval-Valued Datasets

While it is out of the scope of this paper to discuss specific applications of inferential statistics on an interval-valued dataset, we are interested in measuring the amount of information in an interval-valued dataset. Information entropy is the average rate at which information is produced by a stochastic source

of data [24]. Shannon introduced the concept of entropy in his seminal paper “A Mathematical Theory of Communication” [23]. The measure of information entropy associated with each possible data value is:

$$H(x) = - \sum_{i=1}^n p(x_i) \log p(x_i) \quad (15)$$

where $p(x_i)$ is the probability of $x_i \in X$.

An interval-valued dataset $\mathbf{X} = (x_1, x_2, \dots, x_n)$ divides the real axis into $2n + 1$ sub-intervals. Using $\mathcal{P}$ to denote the partition and $x^{(j)}$ to specify its j -th element, we have $\mathcal{P} = (x^{(1)}, x^{(2)}, \dots, x^{(2n+1)})$. As illustrated in Example 2, we can apply Algorithm 2 to find the pdf_j for each $x^{(j)} \in \mathcal{P}$. Then, the probability of $x^{(j)} = \int_{x^{(j)}} pdf_j(t)dt$ is available. Hence, we can apply (15) to calculate the entropy of an interval-valued dataset $\mathbf{X}$. For reader's convenience, we summarize the steps of finding the entropy of $\mathbf{X}$ as an algorithm below.

Algorithm 3: (Finding the entropy for an interval-valued dataset $\mathbf{X}$)

Input: an n -item interval dataset $\mathbf{X}$

pdf_i for all $x_i \in \mathbf{X}$

Output: Entropy($\mathbf{X}$)

Find the partition for the real axis:

Concatenating $\underline{x}$ and $\bar{x}$ as a list c

Sort c

The c forms a $2n + 1$ partition $\mathcal{P}$ of $(-\infty, \infty)$

Find the probability for each $x^{(j)} \in \mathcal{P}$:

For j from 1 to $2n + 1$

Find a pdf_j on $x^{(j)}$ with Algorithm 2

Calculate $p_j = \int_{x^{(j)}} pdf_j(x)dx$

End for

Calculate the entropy:

Entropy($\mathbf{X}$) = 0

For j from 1 to $2n + 1$

Entropy($\mathbf{X}$) = Entropy($\mathbf{X}$) + $p_j \log p_j$

End for

Return Entropy($\mathbf{X}$)

The example below finds the entropy of the sample dataset $\mathbf{X}_0$ with the same assumption of uniform distribution in Example 2.

Example 3. Equation (9) in Example 2 provides the pdf of $\mathbf{X}_0$. Applying it, we obtain the probability of each interval $x^{(j)}$ as

$$p(x) = \int_{x^{(j)}} pdf(t)dt = \begin{cases} 0.025, & x^{(1)} = [1, 1.5] \\ 0.075, & x^{(2)} = [1.5, 2] \\ 0.175, & x^{(3)} = [2, 2.5] \\ 0.197, & x^{(4)} = [2.5, 3] \\ 0.098, & x^{(5)} = [3, 3.5] \\ 0.048, & x^{(6)} = [3.5, 4] \\ 0.194, & x^{(7)} = [4, 5] \\ 0.144, & x^{(8)} = [5, 6] \\ 0.044, & x^{(9)} = [6, 7] \\ 0, & otherwise \end{cases} \quad (16)$$

The entropy of $\mathbf{X}_0$ is $Entropy(X_0) = -\sum_i p_i \log p_i = 2.019$. $\square$

Algorithm 3 provides us a much needed tool in studying point-valued information entropy of an interval-valued dataset. Applying it, we have investigated entropies of the real world financial dataset, which has used in the study of stock market forecasts [6, 7], and [9], from the perspective of information theory. The results are reported in [12]. It not only reveals the deep reason of the significant quality improvements reported before, but also validates the concepts and algorithms presented here in this paper as a successful application.

5 Summary and Future Work

Recent advances have shown that using interval-valued data can significantly improve the quality and efficiency of information processing and uncertainty management. For interval-valued datasets, this work establishes much needed concepts of point-valued variational statistics, probability, and entropy for interval-valued datasets. Furthermore, this paper contains practical algorithms to find these point-valued measures. It provides additional theoretic foundations of applying point-valued methods in analyzing interval-valued datasets.

These point-valued measures enable us to directly apply currently available powerful point-valued statistic, probabilistic, theoretic results to interval-valued datasets. Applying these measures in various applications is definitely among a high priority of our future work. In fact, using this work as the theoretic foundation, we have successfully analyzed the entropies of the real world financial dataset related to the stock market forecasting mentioned in the introduction of this paper. The obtained results are reported in [12] and published in the same volume as this one. On a theoretic side, future work includes extending the concepts in this paper from single dimensional to multi-dimensional interval-valued datasets.

References

1. Bentkowska, U.: New types of aggregation functions for interval-valued fuzzy setting and preservation of pos-B and nec-B-transitivity in decision making problems. *Inf. Sci.* **424**(C), 385–399 (2018)

2. Billard, L., Diday, E.: Regression analysis for interval-valued data. In: Kiers, H.A.L., Rasson, J.P., Groenen, P.J.F., Schader, M. (eds.) *Data Analysis, Classification, and Related Methods. STUDIES CLASS.* Springer, Heidelberg (2000). https://doi.org/10.1007/978-3-642-59789-3_58
3. Dai, J., Wang, W., Mi, J.: Uncertainty measurement for interval-valued information systems. *Inf. Sci.* **251**, 63–78 (2013)
4. Gioia, F., Lauro, C.: Basic statistical methods for interval data. *Statistica Applicata* **17**(1), 75–104 (2005)
5. Grabisch, M., Marichal, J., Mesiar, R., Pap, E.: *Aggregation Functions.* Cambridge University Press, New York (2009)
6. He, L., Hu, C.: Midpoint method and accuracy of variability forecasting. *J. Empir. Econ.* **38**, 705–715 (2009). <https://doi.org/10.1007/s00181-009-0286-6>
7. He, L., Hu, C.: Impacts of interval computing on stock market forecasting. *J. Comput. Econ.* **33**(3), 263–276 (2009). <https://doi.org/10.1007/s10614-008-9159-x>
8. Hu, C., et al.: *Knowledge Processing with Interval and Soft Computing.* Springer, London (2008). <https://doi.org/10.1007/978-1-84800-326-2>
9. Hu, C., He, L.: An application of interval methods to stock market forecasting. *J. Reliable Comput.* **13**, 423–434 (2007). <https://doi.org/10.1007/s11155-007-9039-4>
10. Hu, C.: Using interval function approximation to estimate uncertainty. In: *Interval/Probabilistic Uncertainty and Non-Classical Logics*, pp. 341–352 (2008). https://doi.org/10.1007/978-3-540-77664-2_26
11. Hu, C.: A note on probabilistic confidence of the stock market ILS interval forecasts. *J. Risk Finance* **11**(4), 410–415 (2010)
12. Hu, C., and Hu, Z.: A computational study on the entropy of interval-valued datasets from the stock market. In: Lesot, M.-J., et al. (eds.) *The Proceedings of the 18th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2020)*, IPMU 2020, CCIS, vol. 1239, pp. 422–435. Springer (2020)
13. Huynh, V., Nakamori, Y., Hu, C., Kreinovich, V.: On decision making under interval uncertainty: a new justification of Hurwicz optimism-pessimism approach and its use in group decision making. In: *39th International Symposium on Multiple-Valued Logic*, pp. 214–220 (2009)
14. IEEE Standard for Interval Arithmetic. IEEE Standards Association (2015). <https://standards.ieee.org/standard/1788-2015.html>
15. IEEE Standard for Interval Arithmetic (Simplified). IEEE Standards Association (2018). <https://standards.ieee.org/standard/1788.1-2017.html>
16. de Korvin, A., Hu, C., Chen, P.: Generating and applying rules for interval valued fuzzy observations. In: Yang, Z.R., Yin, H., Everson, R.M. (eds.) *IDEAL 2004.* LNCS, vol. 3177, pp. 279–284. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-28651-6_41
17. Lodwick, W.-A., Jamison, K.-D.: Interval-valued probability in the analysis of problems containing a mixture of possibilistic, probabilistic, and interval uncertainty. *Fuzzy Sets Syst.* **159**(21), 2845–2858 (2008)
18. Moore, R.E.: *Methods and Applications of Interval Analysis.* SIAM Studies in Applied Mathematics, Philadelphia (1979)
19. Marupally, P., Paruchuri, V., Hu, C.: Bandwidth variability prediction with rolling interval least squares (RILS). In: *Proceedings of the 50th ACM SE Conference, Tuscaloosa, AL, USA, 29–31 March 2012*, pp. 209–213. ACM (2012). <https://doi.org/10.1145/2184512.2184562>

20. Nordin, B., Hu, C., Chen, B., Sheng, V.S.: Interval-valued centroids in K-means algorithms. In: Proceedings of the 11th IEEE International Conference on Machine Learning and Applications (ICMLA), Boca Raton, FL, USA, pp. 478–481. IEEE (2012). <https://doi.org/10.1109/ICMLA.2012.87>
21. Pkala, B.: Uncertainty Data in Interval-Valued Fuzzy Set Theory: Properties, Algorithms and Applications, 1st edn. Springer, Cham (2018). <https://doi.org/10.1007/978-3-319-93910-0>
22. Rhodes, C., Lemon, J., Hu, C.: An interval-radial algorithm for hierarchical clustering analysis. In: 14th IEEE International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, pp. 849–856. IEEE (2015)
23. Shannon, C.-E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)
24. Wikipedia: Information entropy. [https://en.wikipedia.org/wiki/Entropy_\(information_theory\)](https://en.wikipedia.org/wiki/Entropy_(information_theory))
25. Wolfram Mathworld. Binary normal distribution. <http://mathworld.wolfram.com/BivariateNormalDistribution.html>