Towards Combining Probabilistic and Interval Uncertainty in Engineering Calculations

S. A. Starks, V. Kreinovich, L. Longpré, M. Ceberio, G. Xiang,
R. Araiza, J. Beck, R. Kandathi, A. Nayak and R. Torres
*NASA Pan-American Center for Earth and Environmental Studies (PACES), University of Texas, El Paso, TX 79968, USA (*vladik@cs.utep.edu*)*

Abstract. In many engineering applications, we have to combine probabilistic and interval errors. For example, in environmental analysis, we observe a pollution level $x(t)$ in a lake at different moments of time t, and we would like to estimate standard statistical characteristics such as mean, variance, autocorrelation, correlation with other measurements. In environmental measurements, we often only know the values with interval uncertainty. We must therefore modify the existing statistical algorithms to process such interval data. Such modification are described in this paper.

Keywords: probabilistic uncertainty, interval uncertainty, engineering calculations

1. Formulation of the Problem

Computing statistics is important. In many engineering applications, we are interested in computing statistics. For example, in environmental analysis, we observe a pollution level $x(t)$ in a lake at different moments of time t, and we would like to estimate standard statistical characteristics such as mean, variance, autocorrelation, correlation with other measurements. For each of these characteristics C, there is an expression $C(x_1, \ldots, x_n)$ that enables us to provide an estimate for C based on the observed values $x_1, \ldots, x_n$. For example, a reasonable statistic for estimating the mean value of a probability distribution is the population average $E(x_1, \ldots, x_n) = \dfrac{1}{n}(x_1 + \ldots + x_n)$; a reasonable statistic for estimating the variance V is the population variance $V(x_1, \ldots, x_n) = \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} (x_i - \bar{x})^2$, where $\bar{x} \stackrel{\text{def}}{=} \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} x_i$.

Interval uncertainty. In environmental measurements, we often only know the values with interval uncertainty. For example, if we did not detect any pollution, the pollution value v can be anywhere between 0 and the sensor's detection limit DL. In other words, the only information that we have about v is that v belongs to the interval $[0, DL]$; we

have no information about the probability of different values from this interval.

Another example: to study the effect of a pollutant on the fish, we check on the fish daily; if a fish was alive on Day 5 but dead on Day 6, then the only information about the lifetime of this fish is that it is somewhere within the interval $[5, 6]$; we have no information about the probability of different values within this interval.

In non-destructive testing, we look for outliers as indications of possible faults. To detect an outlier, we must know the mean and standard deviation of the normal values – and these values can often only be measured with interval uncertainty (see, e.g., (Rabinovich, 1993; Osegueda et al., 2002)). In other words, often, we know the result $\widetilde{x}$ of measuring the desired characteristic x, and we know the upper bound Δ on the absolute value $|\Delta x|$ of the measurement error $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$ (this upper bound is provided by the manufacturer of the measuring instrument), but we have no information about the probability of different values $\Delta x \in [-\Delta, \Delta]$. In such situations, after the measurement, the only information that we have about the actual value x of the measured quantity is that this value belongs to interval $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$.

In geophysics, outliers should be identified as possible locations of minerals; the importance of interval uncertainty for such applications was emphasized in (Nivlet et al., 2001; Nivlet et al., 2001a). Detecting outliers is also important in bioinformatics (Shmulevich and Zhang, 2002).

In bioinformatics and bioengineering applications, we must solve systems of linear equations in which coefficients come from experts and are only known with interval uncertainty; see, e.g., (Zhang et al., 2004).

In biomedical systems, statistical analysis of the data often leads to improvements in medical recommendations; however, to maintain privacy, we do not want to use the exact values of the patient's parameters. Instead, for each parameter, we select fixed values, and for each patient, we only keep the corresponding range. For example, instead of keeping the exact age, we only record whether the age is between 0 and 10, 10 and 20, 20 and 30, etc. We must then perform statistical analysis based on such interval data; see, e.g., (Kreinovich and Longpré, 2003; Xiang et al., 2004).

Estimating statistics under interval uncertainty: a problem. In all such cases, instead of the actual values $x_1, \ldots, x_n$, we only know the intervals $\mathbf{x}_1 = [\underline{x}_1, \overline{x}_1], \ldots, \mathbf{x}_n = [\underline{x}_n, \overline{x}_n]$ that contain the (unknown) actual values of the measured quantities. For different values $x_i \in \mathbf{x}_i$, we get, in general, different values of the corresponding statistical characteristic $C(x_1, \ldots, x_n)$. Since all values $x_i \in \mathbf{x}_i$ are possible, we conclude that

all the values $C(x_1, \ldots, x_n)$ corresponding to $x_i \in \mathbf{x}_i$ are possible estimates for the corresponding statistical characteristic. Therefore, for the interval data $\mathbf{x}_1, \ldots, \mathbf{x}_n$, a reasonable estimate for the corresponding statistical characteristic is the range

$$C(\mathbf{x}_1, \ldots, \mathbf{x}_n) \stackrel{\text{def}}{=} \{C(x_1, \ldots, x_n) \,|\, x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}.$$

We must therefore modify the existing statistical algorithms so that they would be able to estimate such ranges. This is a problem that we solve in this paper.

This problem is a part of a general problem. The above range estimation problem is a specific problem related to a combination of interval and probabilistic uncertainty. Such problems – and their potential applications – have been described, in a general context, in the monographs (Kuznetsov, 1991; Walley, 1991); for further developments, see, e.g., (Rowe, 1988; Williamson, 1990; Berleant, 1993; Berleant, 1996; Berleant and Goodman-Strauss, 1998; Ferson et al., 2001; Ferson, 2002; Berleant et al., 2003; Lodwick and Jamison, 2003; Moore and Lodwick, 2003; Regan et al., (in press)) and references therein.

2. Analysis of the Problem

Mean. Let us start our discussion with the simplest possible characteristic: the mean. The arithmetic average E is a monotonically increasing function of each of its n variables $x_1, \ldots, x_n$, so its smallest possible value $\underline{E}$ is attained when each value x_i is the smallest possible ($x_i = \underline{x}_i$) and its largest possible value is attained when $x_i = \overline{x}_i$ for all i. In other words, the range $\mathbf{E}$ of E is equal to $[E(\underline{x}_1, \ldots, x_n), E(\overline{x}_1, \ldots, \overline{x}_n)]$. In other words, $\underline{E} = \dfrac{1}{n}(\underline{x}_1 + \ldots + \underline{x}_n)$ and $\overline{E} = \dfrac{1}{n}(\overline{x}_1 + \ldots + \overline{x}_n)$.

Variance: computing the exact range is difficult. Another widely used statistic is the variance. In contrast to the mean, the dependence of the variance V on x_i is not monotonic, so the above simple idea does not work. Rather surprisingly, it turns out that the problem of computing the exact range for the variance over interval data is, in general, NP-hard (Ferson et al., 2002; Kreinovich, (in press)) which means, crudely speaking, that the worst-case computation time grows exponentially with n. Moreover, if we want to compute the variance range with a given accuracy ε, the problem is still NP-hard. (For a more detailed

description of NP-hardness in relation to interval uncertainty, see, e.g., (Kreinovich et al., 1997).)

Linearization. From the practical viewpoint, often, we may not need the exact range, we can often use approximate linearization techniques. For example, when the uncertainty comes from measurement errors Δx_i, and these errors are small, we can ignore terms that are quadratic (and of higher order) in Δx_i and get reasonable estimates for the corresponding statistical characteristics. In general, in order to estimate the range of the statistic $C(x_1, \ldots, x_n)$ on the intervals $[\underline{x}_1, \overline{x}_1]$, $\ldots$, $[\underline{x}_n, \overline{x}_n]$, we expand the function C in Taylor series at the midpoint $\widetilde{x}_i \stackrel{\text{def}}{=} (\underline{x}_i + \overline{x}_i)/2$ and keep only linear terms in this expansion. As a result, we replace the original statistic with its linearized version $C_{\text{lin}}(x_1, \ldots, x_n) = C_0 - \sum_{i=1}^{n} C_i \cdot \Delta x_i$, where $C_0 \stackrel{\text{def}}{=} C(\widetilde{x}_1, \ldots, \widetilde{x}_n)$, $C_i \stackrel{\text{def}}{=} \dfrac{\partial C}{\partial x_i}(\widetilde{x}_1, \ldots, \widetilde{x}_n)$, and $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$. For each i, when $x_i \in [\underline{x}_i, \overline{x}_i]$, the difference Δx_i can take all possible values from $-\Delta_i$ to Δ_i, where $\Delta_i \stackrel{\text{def}}{=} (\overline{x}_i - \underline{x}_i)/2$. Thus, in the linear approximation, we can estimate the range of the characteristic C as $[C_0 - \Delta, C_0 + \Delta]$, where $\Delta \stackrel{\text{def}}{=} \sum_{i=1}^{n} |c_i| \cdot \Delta_i$.

In particular, for variance, $C_i = \dfrac{\partial V}{\partial x_i} = \dfrac{2}{n}(\widetilde{x}_i - \bar{\widetilde{x}})$, where $\bar{\widetilde{x}}$ is the average of the midpoints $\widetilde{x}_i$. So, here, $V_0 = \dfrac{1}{n} \sum_{i=1}^{n} (\widetilde{x}_i - \bar{\widetilde{x}})^2$ is the variance of the midpoint values $\widetilde{x}_1, \ldots, \widetilde{x}_n$, and $\Delta = \dfrac{2}{n} \sum_{i=1}^{n} |\widetilde{x}_i - \bar{\widetilde{x}}| \cdot \Delta_i$.

It is worth mentioning that for the variance, the ignored quadratic term is equal to $\dfrac{1}{n} \sum_{i=1}^{n} (\Delta x_i)^2 - (\overline{\Delta x})^2$, where $\overline{\Delta x} \stackrel{\text{def}}{=} \dfrac{1}{n} \sum_{i=1}^{n} \Delta x_i$, and therefore, can be bounded by 0 from below and by $\Delta^{(2)} \stackrel{\text{def}}{=} \dfrac{1}{n} \sum_{i=1}^{n} \Delta_i^2$ from above. Thus, the interval $[V_0 - \Delta, V_0 + \Delta + \Delta^{(2)}]$ is a guaranteed enclosure for $\mathbf{V}$.

Linearization is not always acceptable. In some cases, linearized estimates are not sufficient: the intervals may be wide so that quadratic terms can no longer be ignored, and/or we may be in a situation where we want to guarantee that, e.g., the variance does not exceed a certain required threshold. In such situations, we need to get the exact range – or at least an enclosure for the exact range.

Since, even for as simple a characteristic as variance, the problem of computing its exact range is NP-hard, we cannot have a feasible-time algorithm that always computes the exact range of these characteristics. Therefore, we must look for the reasonable classes of problems for which such algorithms are possible. Let us analyze what such classes can be.

First class: narrow intervals. As we have just mentioned, the computational problems become more complex when we have wider intervals. In other words, when intervals are narrower, the problems are easier. How can we formalize "narrow intervals"? One way to do it is as follows: the actual values $x_1, \ldots, x_n$ of the measured quantity are real numbers, so they are usually different. The data intervals $\mathbf{x}_i$ contain these values. When the intervals $\mathbf{x}_i$ surrounding the corresponding points x_i are narrow, these intervals do not intersect. When their widths becomes larger than the distance between the original values, the intervals start intersecting.

Definition. Thus, the ideal case of "narrow intervals" can be described as the case when no two intervals $\mathbf{x}_i$ intersect.

Second class: slightly wider intervals. Slightly wider intervals correspond to the situation when few intervals intersect, i.e., when for some integer K, no set of K intervals has a common intersection.

Third class: single measuring instrument. Since we want to find the exact range $\mathbf{C}$ of a statistic C, it is important not only that intervals are relatively narrow, it is also important that they are approximately of the same size: otherwise, if, say, Δx_i^2 is of the same order as Δx_j, we cannot meaningfully ignore Δx_i^2 and retain Δx_j. In other words, the interval data set should not combine high-accurate measurement results (with narrow intervals) and low-accurate results (with wide intervals): all measurements should have been done by a single measuring instrument (or at least by several measuring instruments of the same type).

How can we describe this mathematically? A clear indication that we have two measuring instruments (MI) of different quality is that one interval is a proper subset of the other one: $[\underline{x}_i, \overline{x}_i] \subseteq (\underline{x}_j, \overline{x}_j)$.

Definition. So, if all pairs of non-degenerate intervals satisfy the following *subset property* $[\underline{x}_i, \overline{x}_i] \not\subseteq (\underline{x}_j, \overline{x}_j)$, we say that the measurements were done *by a single MI.*

Comment. This restriction only refers to inexact measurement results, i.e., to non-degenerate intervals. In additional to such interval

values, we may have exact values (degenerate intervals). For example, in geodetic measurements, we may select some point ("benchmark") as a reference point, and describe, e.g., elevation of each point relative to this benchmark. For the benchmark point itself, the relative elevation will be therefore exactly equal to 0. When we want to compute the variance of elevations, we want to include the benchmark point too. From this viewpoint, when we talk about measurements made by a single measuring instrument, we may allow degenerate intervals (i.e., exact numbers) as well.

A reader should be warned that in the published algorithms describing a single MI case (Xiang et al., 2004), we only considered non-degenerate intervals. However, as one can easily see from the published proofs (and from the idea of these proofs, as described below), these algorithms can be easily modified to incorporate possible exact values x_i.

Fourth class: same accuracy measurement. In some situations, it is also reasonable to consider a specific case of the single MI case when all measurements are performed with exactly the same accuracy, i.e., in mathematical terms, when all non-degenerate intervals $[\underline{x}_i, \overline{x}_i]$ have exactly the same half-width $\Delta_i = \dfrac{1}{2} \cdot (\overline{x}_i - \underline{x}_i)$.

Fifth class: several MI. After the single MI case, the natural next case is when we have several MI, i.e., when our intervals are divided into several subgroups each of which has the above-described subset property.

Sixth class: privacy case. Although these definitions are in terms of measurements, they make sense for other sources of interval data as well. For example, for privacy data, intervals either coincide (if the value corresponding to the two patients belongs to the same range) or are different, in which case they can only intersect in one point. Similarly to the above situation, we also allow exact values in addition to ranges; these values correspond, e.g., to the exact records made in the past, records that are already in the public domain.

Definition. We will call interval data with this property – that every two non-degenerate intervals either coincide or do nor intersect – *privacy case.*

Comment. For the privacy case, the subset property is satisfied, so algorithms that work for a single MI case work for the privacy case as well.

Seventh class: non-detects. Similarly, if the only source of interval uncertainty is detection limits, i.e., if every measurement result is either an exact value or a *non-detect*, i.e., an interval $[0, DL_i]$ for some real number DL_i (with possibly different detection limits for different sensors), then the resulting non-degenerate intervals also satisfy the subset property. Thus, algorithms that work for a single MI case work for this "non-detects" case as well.

Also, an algorithm that works for the general privacy case also works for the non-detects case when all sensors have the same detection limit DL.

3. Results

Variance: known results. The lower bound $\underline{V}$ can be always computed in time $O(n \cdot \log(n))$ (Granvilliers et al., 2004).

Computing $\overline{V}$ is, in general, an NP-hard problem; $\overline{V}$ can be computed in time 2^n. If intervals do not intersect (and even if "narrowed" intervals $[\widetilde{x}_i - \Delta_i/n, \widetilde{x}_i + \Delta_i/n]$ do not intersect), we can compute $\overline{V}$ in time $O(n \cdot \log(n))$ (Granvilliers et al., 2004). If for some K, no more than K interval intersect, we can compute $\overline{V}$ in time $O(n^2)$ (Ferson et al., 2002; Kreinovich, (in press)).

For the case of a single MI, $\overline{V}$ can be computed in time $O(n \cdot \log(n))$; for m MIs, we need time $O(n^{m+1})$ (Xiang et al., 2004).

Variance: main ideas behind the known results. The algorithm for computing $\underline{V}$ is based on the fact that when a function V attains a minimum on an interval $[\underline{x}_i, \overline{x}_i]$, then either $\dfrac{\partial V}{\partial x_i} = 0$, or the minimum is attained at the left endpoint $x_i = \underline{x}_i$ – then $\dfrac{\partial V}{\partial x_i} > 0$, or $x_i = \overline{x}_i$ and $\dfrac{\partial V}{\partial x_i} < 0$. Since the partial derivative is equal to $(2/n) \cdot (x_i - \bar{x})$, we conclude that either $x_i = \bar{x}$, or $x_i = \underline{x}_i > \bar{x}$, or $x_i = \underline{x}_i < \bar{x}$. Thus, if we know where $\bar{x}$ is located in relation to all the endpoints, we can uniquely determine the corresponding minimizing value x_i for every i: if $\overline{x}_i \le \bar{x}$ then $x_i = \overline{x}_i$; if $x_i \le \underline{x}_i$, then $x_i = \underline{x}_i$; otherwise, $x_i = \bar{x}$. The corresponding value $\bar{x}$ can be found from the condition that $\bar{x}$ is the average of all the selected values x_i.

So, to find the smallest value of V, we can sort all $2n$ bounds $\underline{x}_i, \overline{x}_i$ into a sequence $x_{(1)} \le x_{(2)} \le \ldots$; then, for each zone $[x_{(k)}, x_{(k+1)}]$, we compute the corresponding values x_i, find their variance V_k, and then compute the smallest of these variances V_k.

For each of $2n$ zones, we need $O(n)$ steps, so this algorithm requires $O(n^2)$ steps. It turns out that the function V_k decreases until the desired k then increases, so we can use binary search – that requires that we only analyze $O(\log(n))$ zones – find the appropriate zone k. As a result, we get an $O(n \cdot \log(n))$ algorithm.

For $\overline{V}$, to the similar analysis of the derivatives, we can add the fact that the second derivative of V is ≥ 0, so there cannot be a maximum inside the interval $[\underline{x}_i, \overline{x}_i]$. So, in principle, to compute $\overline{V}$, it is sufficient to consider all 2^n combinations of endpoints. When few intervals intersect, then, when $\overline{x}_i \leq \bar{x}$, we take $x_i = \underline{x}_i$; when $\bar{x} \leq \underline{x}_i$, we take $x_i = \overline{x}_i$; otherwise, we must consider both possibilities $x_i = \underline{x}_i$ and $x_i = \overline{x}_i$.

For the case of a single MI, we can sort the intervals in lexicographic order: $\mathbf{x}_i \leq \mathbf{x}_j$ if and only if $\underline{x}_i < \underline{x}_j$ or ($\underline{x}_i = \underline{x}_j$ and $\overline{x}_i \leq \overline{x}_j$). It can be proven that the maximum of V is always attained if for some k, the first k values x_i are equal to $\underline{x}_i$ and the next $n - k$ values x_i are equal to $\overline{x}_i$. This result is proven by reduction to a contradiction: if in the maximizing vector $x = (x_1, \ldots, x_n)$, some $\overline{x}_i$ is preceding some $\underline{x}_j$, $i < j$, then we can increase V while keeping E intact – which is in contradiction with the assumption that the vector x was maximizing. Specifically, to increase V, we can do the following: if $\Delta_i \leq \Delta_j$, we replace $\overline{x}_i$ with $\underline{x}_i = \overline{x}_i - 2\Delta_i$ and $\underline{x}_j$ with $\underline{x}_j + 2\Delta_i$; otherwise, we replace $\underline{x}_j$ with $\overline{x}_j = \underline{x}_j + 2\Delta_j$ and $\overline{x}_i$ with $\overline{x}_i - 2\Delta_j$.

As a result, to find the maximum of V, it is sufficient to sort the intervals (this takes $O(n \cdot \log(n))$ time), and then, for different values k, check vectors $(\underline{x}_1, \ldots, \underline{x}_k, \overline{x}_{k+1}, \ldots, \overline{x}_n)$. The dependence of V on k is concave, so we can use binary search to find k; binary search takes $O(\log(n))$ steps, and for each k, we need linear time, so overall, we need time $O(n \cdot \log(n))$.

In case of several MI, we sort intervals corresponding to each of m MI. Then, to find the maximum of V, we must find the values $k_1, \ldots, k_m$ corresponding to m MIs. There are $\leq n^m$ combinations of k_is, and checking each combination requires $O(n)$ time, so overall, we need time $O(n^{m+1})$.

Variance: new results. Sometimes, most of the data is accurate, so among n intervals, only $d \ll n$ are non-degenerate intervals. For example, we can have many accurate values and m non-detects. In this situation, to find the extrema of V, we only need to find x_i for d non-degenerate intervals; thus, we only need to consider $2d$ zones formed by their endpoints. Within each zone, we still need $O(n)$ computations to compute the corresponding variance.

So, in this case, to compute $\underline{V}$, we need time $O(n \cdot \log(d))$, and to compute $\overline{V}$, we need $O(n \cdot 2^d)$ steps. If narrowed intervals do not intersect, we need time $O(n \cdot \log(d))$ to compute $\overline{V}$; if for some K, no more than K interval intersect, we can compute $\overline{V}$ in time $O(n \cdot d)$.

For the case of a single MI, $\overline{V}$ can be computed in time $O(n \cdot \log(d))$; for m MIs, we need time $O(n \cdot d^m)$.

In addition to new algorithms, we also have a new NP-hardness result. In the original proof of NP-hardness, we have $\widetilde{x}_1 = \ldots = \widetilde{x}_n = 0$, i.e., all measurement results are the same, only accuracies Δ_i are different. What if all the measurement results are different? We can show that in this case, computing $\overline{V}$ is still an NP-hard problem: namely, for every n-tuple of real numbers $\widetilde{x}_1, \ldots, \widetilde{x}_n$, the problem of computing $\overline{V}$ for intervals $\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$ is still NP-hard.

To prove this result, it is sufficient to consider $\Delta_i = N \cdot \Delta_i^{(0)}$, where $\Delta_i^{(0)}$ are the values used in the original proof. In this case, we can describe $\Delta x_i = \widetilde{x}_i - x_i$ as $N \cdot \Delta x_i^{(0)}$, where $\Delta_i^{(0)} \in [-\Delta_i^{(0)}, \Delta_i^{(0)}]$. For large N, the difference between the variance corresponding to the values $x_i = \widetilde{x}_i + N \cdot \Delta x_i^{(0)}$ and N^2 times the variance of the values $\Delta x_i^{(0)}$ is bounded by a term proportional to N (and the coefficient at N can be easily bounded). Thus, the difference between $\overline{V}$ and $N^2 \cdot \overline{V}^{(0)}$ is bounded by $C \cdot N$ for some known constant C. Hence, by computing $\overline{V}$ for sufficiently large N, we can compute $\overline{V}^{(0)}$ with a given accuracy $\varepsilon > 0$, and we already know that computing $\overline{V}^{(0)}$ with given accuracy is NP-hard. This reduction proves that our new problem is also NP-hard.

Covariance: known results. In general, computing the range of covariance $C_{xy} = \dfrac{1}{n} \sum\limits_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$ based on given intervals $\mathbf{x}_i$ and $\mathbf{y}_i$ is NP-hard (Osegueda et al., 2002). When boxes $\mathbf{x}_i \times \mathbf{y}_i$ do not intersect – or if $\geq K$ boxes cannot have a common point – we can compute the range in time $O(n^3)$ (Beck et al., 2004).

The main idea behind this algorithm is to consider the derivatives of C relative to x_i and y_i. Then, once we know where the point $(\bar{x}, \bar{y})$ is in relation to x_i and y_i, we can uniquely determine the optimizing values x_i and y_i – except for the boxes $\mathbf{x}_i \times \mathbf{y}_i$ that contain $(\bar{x}, \bar{y})$. The bounds $\underline{x}_i$ and $\overline{x}_i$ divide the x axis into $2n + 2$ intervals; similarly, the y-bounds divide the y-axis into $2n + 2$ intervals. Combining these intervals, we get $O(n^2)$ zones. Due to the limited intersection property, for each of these zones, we have finitely many ($\leq K$) indices i for which the corresponding box intersects with the zone. For each such box, we may have two different combinations: $(\underline{x}_i, \underline{y}_i)$ and $(\overline{x}_i, \overline{y}_i)$ for $\overline{C}$ and

$(\underline{x}_i, \overline{y}_i)$ and $(\overline{x}_i, \underline{y}_i)$ for $\underline{C}$. Thus, we have finitely many ($\leq 2^K$) possible combinations of (x_i, y_i) corresponding to each zone. Hence, for each of $O(n^2)$ zones, it takes $O(n)$ time to find the corresponding values x_i and y_i and to compute the covariance; thus, overall, we need $O(n^3)$ time.

Covariance: new results. If $n-d$ measurement results (x_i, y_i) are exact numbers and only d are non-point boxes, then we only need $O(d^2)$ zones, so we can compute the range in time $O(n \cdot d^2)$.

In the privacy case, all boxes $\mathbf{x}_i \times \mathbf{y}_i$ are either identical or non-intersecting, so the only case when a box intersects with a zone is when the box coincides with this zone. For each zone k, there may be many (n_k) such boxes, but since they are all identical, what matters for our estimates is how many of them are assigned one of the possible (x_i, y_i) combinations and how many the other one. There are only $n_k + 1$ such assignments: 0 to first combination and n_k to second, 1 to first and $n_k - 1$ to second, etc. Thus, the overall number of all combinations for all the zones k is $\sum\limits_k n_k + \sum\limits_k 1$, where $\sum n_k = n$ and $\sum\limits_k 1$ is the overall number of zones, i.e., $O(n^2)$. For each combination of x_i and y_i, we need $O(n)$ steps. Thus, in the privacy case, we can compute both $\underline{C}$ and $\overline{C}$ in time $O(n^2) \cdot O(n) = O(n^3)$ (or $O(n \cdot d^2)$ if only d boxes are non-degenerate).

Another polynomial-time case is when all the measurements are exactly of the same accuracy, i.e., when all non-degenerate x-intervals have the same half-width Δ_x, and all non-degenerate y-intervals have the same half-width Δ_y. In this case, e.g., for $\overline{C}$, if we have at least two boxes i and j intersecting with the same zone, and we have $(x_i, y_i) = (\underline{x}_i, \underline{y}_i)$ and $(x_j, y_j) = (\overline{x}_j, \overline{y}_j)$, then we can swap i and j assignments – i.e., make $(x_i', y_i') = (\overline{x}_i, \overline{y}_i)$ and $(x_j', y_j') = (\underline{x}_j, \underline{y}_j)$ – without changing $\bar{x}$ and $\bar{y}$. In this case, the only change in C_{xy} comes from replacing $x_i \cdot y_i + x_j \cdot y_j$. It is easy to see that the new value C is larger than the old value if and only if $z_i > z_j$, where $z_i \stackrel{\text{def}}{=} \widetilde{x}_i \cdot \Delta_y + \widetilde{y}_i \cdot \Delta_x$. Thus, in the true maximum, whenever we assign $(\underline{x}_i, \underline{y}_i)$ to some i and $(\overline{x}_i, \overline{y}_j)$ to some j, we must have $z_i \leq z_j$. So, to get the largest value of C, we must sort the indices by z_i, select a threshold t, and assign $(\underline{x}_i, \underline{y}_i)$ to all the boxes with $z_i \leq t$ and $(\overline{x}_j, \overline{y}_j)$ to all the boxes j with $z_j > t$. If $n_k \leq n$ denotes the overall number of all the boxes that intersect with k-th zone, then we have $n_k + 1$ possible choices of thresholds, hence $n_k + 1$ such assignments. For each of $O(n^2)$ zones, we test $\leq n$ assignments; testing each assignment requires $O(n)$ steps, so overall, we need time $O(n^4)$.

If only d boxes are non-degenerate, we only need time $O(n \cdot d^3)$.

Detecting outliers: known results. Traditionally, in statistics, we fix a value k_0 (e.g., 2 or 3) and claim that every value x outside the k_0-sigma interval $[L, U]$, where $L \stackrel{\text{def}}{=} E - k_0 \cdot \sigma$, $U \stackrel{\text{def}}{=} E + k_0 \cdot \sigma$ (and $\sigma \stackrel{\text{def}}{=} \sqrt{V}$), is an outlier; thus, to detect outliers based on interval data, we must know the ranges of L and U. It turns out that we can always compute $\underline{U}$ and $\overline{L}$ in $O(n^2)$ time (Kreinovich et al., 2003a; Kreinovich et al., 2004). In contrast, computing $\overline{U}$ and $\underline{L}$ is NP-hard; in general, it can be done in 2^n time, and in quadratic time if $\leq K$ intervals intersect (even if $\leq K$ appropriately narrowed intervals intersect) (Kreinovich et al., 2003a; Kreinovich et al., 2004).

For every x, we can also determine the "degree of outlier-ness" R as the smallest k_0 for which $x \notin [E - k_0 \cdot \sigma, E + k_0 \cdot \sigma]$, i.e., as $|x - E|/\sigma$. It turns out that $\overline{R}$ can be always computed in time $O(n^2)$; the lower bound $\underline{R}$ can be also computed in quadratic time if $\leq K$ narrowed intervals intersect (Kreinovich et al., 2003a).

Detecting outliers: new results. Similar to the case of variance, if we only have $d \ll n$ non-degenerate intervals, then instead of $O(n^2)$ steps, we only need $O(n \cdot d)$ steps (and instead of 2^n steps, we only need $O(n \cdot 2^d)$ steps).

For the case of a single MI, similarly to variance, we can prove that the maximum of U and the minimum of L are attained at one of the vectors $(\underline{x}_1, \ldots, \underline{x}_k, \overline{x}_{k+1}, \ldots, \overline{x}_n)$; actually, practically the same proof works, because increasing V without changing E increases $U = E + k_0 \cdot \sqrt{V}$ as well. Thus, to find $\overline{U}$ and $\underline{L}$, it is sufficient to check n such sequences; checking each sequence requires $O(n)$ steps, so overall, we need $O(n^2)$ time. For m MI, we need $O(n^{m+1})$ time.

If only $d \ll n$ intervals are non-degenerate, then we need, correspondingly, time $O(n \cdot d)$ and $O(n \cdot d^m)$.

Moments. For population moments $\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^q$, known interval bounds on x^q leads to exact range. For central moments $M_q = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^q$, we have the following results (Kreinovich et al., 2004a). For even q, the lower endpoint $\underline{M}_q$ can be computed in $O(n^2)$ time; the upper endpoint $\overline{M}_q$ can always be computed in time $O(2^n)$, and in $O(n^2)$ time if $\leq K$ intersect. For odd q, if $\leq K$ intervals do not intersect, we can compute both $\underline{M}_q$ and $\overline{M}_q$ in $O(n^3)$ time.

If only d out of n intervals are non-degenerate, then we need $O(n \cdot 2^d)$ time instead of $O(2^n)$, $O(n \cdot d)$ instead of $O(n^2)$, and $O(n \cdot d^2)$ instead of $O(n^3)$.

For even q, we can also consider the case of a single MI. The arguments work not only for M_q, but also for a generalized central moment $M_\psi \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \psi(x_i - E)$ for an arbitrary convex function $\psi(x) \geq 0$ for which $\psi(0) = 0$ and $\psi''(x) > 0$ for all $x \neq 0$. Let us first show that the maximum cannot be attained inside an interval $[\underline{x}_i, \overline{x}_i]$. Indeed, in this case, at the maximizing point, the first derivative

$$\frac{\partial M_\psi}{\partial x_i} = \frac{1}{n} \cdot \psi'(x_i - E) - \frac{1}{n^2} \cdot \sum_{j=1}^{n} \psi'(x_j - E)$$

should be equal to 0, and the second derivative

$$\frac{\partial^2 M_\psi}{\partial x_i^2} = \frac{1}{n} \cdot \psi''(x_i - E) \cdot \left(1 - \frac{2}{n}\right) + \frac{1}{n^3} \cdot \sum_{j=1}^{n} \psi''(x_j - E)$$

is non-positive. Since the function $\psi(x)$ is convex, we have $\psi''(x) \geq 0$, so this second derivative is a sum of non-negative terms, and the only case when it is non-negative is when all these terms are 0s, i.e., when $x_j = E$ for all j. In this case, $M_\psi = 0$ which, for non-degenerate intervals, is clearly not the largest possible value of M_ψ.

So, for every i, the maximum of M_ψ is attained either when $x_i = \underline{x}_i$ or when $x_i = \overline{x}_i$. Similarly to the proof for the variance, we will now prove that the maximum is always attained for one of the vectors $(\underline{x}_1, \ldots, \underline{x}_k, \overline{x}_{k+1}, \ldots, \overline{x}_n)$. To prove this, we need to show that if $x_i = \overline{x}_i$ and $x_j = \underline{x}_j$ for some $i < j$ (and $\underline{x}_i \leq \underline{x}_j$), then the change described in that proof, while keeping the average E intact, increases the value of M_ψ. Without losing generality, we can consider the case $\Delta_i \leq \Delta_j$. In this case, the fact that M_ψ increase after the above-described change is equivalent to: $\psi(\underline{x}_i + 2\Delta_i - E) + \psi(\underline{x}_j - E) \leq \psi(\underline{x}_i - E) + \psi(\underline{x}_j + 2\Delta_i - E)$, i.e., that $\psi(\underline{x}_i + 2\Delta_i - E) - \psi(\underline{x}_i - E) \leq \psi(\underline{x}_j + 2\Delta_j - E) - \psi(\underline{x}_j - E)$. Since $\underline{x}_i \leq \underline{x}_j$ and $\underline{x}_i - E \leq \underline{x}_j - E$, this can be proven if we show that for every $\Delta > 0$ (and, in particular, for $\Delta = 2\Delta_i$), the function $\psi(x + \Delta) - \psi(x)$ is increasing. Indeed, the derivative of this function is equal to $\psi'(x + \Delta) - \psi'(x)$, and since $\psi''(x) \geq 0$, we do have $\psi'(x + \Delta) \geq \psi'(x)$.

Therefore, to find $\overline{M}_\psi$, it is sufficient to check all n vectors of the type $(\underline{x}_1, \ldots, \underline{x}_k, \overline{x}_{k+1}, \ldots, \overline{x}_n)$, which requires $O(n^2)$ steps. For m MIs, we similarly need $O(n^{m+1})$ steps.

Summary. These results are summarized in the following table. In this table, the first row corresponds to a general case, other rows correspond to different classes of problems described in Section 2:

class number	class description
0	general case
1	narrow intervals: no intersection
2	slightly wider intervals $\leq K$ intervals intersect
3	single measuring instrument (MI): subset property – no interval is a "proper" subset of the other
4	same accuracy measurements: all intervals have the same half-width
5	several (m) measuring instruments: intervals form m groups, with subset property in each group
6	privacy case: intervals same or non-intersecting
7	non-detects case: only non-degenerate intervals are $[0, DL_i]$

#	E	V	C_{xy}	L, U, R	M_{2p}	M_{2p+1}
0	$O(n)$	NP-hard	NP-hard	NP-hard	NP-hard	?
1	$O(n)$	$O(n \cdot \log(n))$	$O(n^3)$	$O(n^2)$	$O(n^2)$	$O(n^3)$
2	$O(n)$	$O(n^2)$	$O(n^3)$	$O(n^2)$	$O(n^2)$	$O(n^3)$
3	$O(n)$	$O(n \cdot \log(n))$	?	$O(n^2)$	$O(n^2)$	?
4	$O(n)$	$O(n \cdot \log(n))$	$O(n^4)$	$O(n^2)$	$O(n^2)$	?
5	$O(n)$	$O(n^{m+1})$	?	$O(n^{m+1})$	$O(n^{m+1})$	?
6	$O(n)$	$O(n \cdot \log(n))$	$O(n^3)$	$O(n^2)$	$O(n^2)$	?
7	$O(n)$	$O(n \cdot \log(n))$	?	$O(n^2)$	$O(n^2)$	?

The case when only d out of n data points are intervals is summarized in the following table:

#	E	V	C_{xy}	L, U, R	M_{2p}	M_{2p+1}
0	$O(n)$	NP-hard	NP-hard	NP-hard	NP-hard	?
1	$O(n)$	$O(n \log(d))$	$O(n \cdot d^2)$	$O(n \cdot d)$	$O(nd)$	$O(nd^2)$
2	$O(n)$	$O(nd)$	$O(n \cdot d^2)$	$O(n \cdot d)$	$O(nd)$	$O(nd^2)$
3	$O(n)$	$O(n \log(d))$	?	$O(n \cdot d)$	$O(nd)$	?
4	$O(n)$	$O(n \log(d))$	$O(n \cdot d^3)$	$O(n \cdot d)$	$O(nd)$	?
5	$O(n)$	$O(nd^m)$	?	$O(n \cdot d^m)$	$O(nd^m)$	?
6	$O(n)$	$O(n \log(d))$	$O(n \cdot d^2)$	$O(n \cdot d)$	$O(nd)$	?
7	$O(n)$	$O(n \log(d))$	?	$O(n \cdot d)$	$O(nd)$	?

Weighted mean and weighted average. In the above text, we considered the case when we only know the upper bound Δ_i on the overall measurement error. In some real-life situations (see, e.g., (Rabinovich, 1993)), we know the standard deviation σ_i of the random error component and the bound Δ_i on the absolute value of the systematic error component. If we had no systematic errors, then we would able to estimate the mean E by solving the corresponding Least Squares problem $\sum \sigma_i^{-2} \cdot (x_i - E)^2 \to \min_E$, i.e., as $E_w = \sum_{i=1}^{n} p_i \cdot x_i$, where $p_i \stackrel{\text{def}}{=} \dfrac{\sigma_i^{-2}}{\sum_{j=1}^{n} \sigma_j^{-2}}$. In this case, the variance can be estimated as $V_w = \sum_{i=1}^{n} p_i \cdot (x_i - E_w)^2 = \sum_{i=1}^{n} p_i \cdot x_i^2 - E_w^2$. Due to the presence of systematic errors, the actual values x_i may be anywhere within the intervals $[\underline{x}_i, \overline{x}_i] \stackrel{\text{def}}{=} [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$. Thus, we arrive at the problem of estimating the range of the above expressions for weighted mean and weighted variance on the interval data $[\underline{x}_i, \overline{x}_i]$.

The expression for the mean is monotonic, so, similar to the average, we substitute the values $\underline{x}_i$ to get $\underline{E}_w$ and the values $\overline{x}_i$ to get $\overline{E}_w$.

For the weighted variance, the derivative is equal to $2p_i \cdot (x_i - E_w)$, and the second derivative is always ≥ 0, so, similarly to the above

proof for the non-weighted variance, we conclude that the minimum is always attained at a vector $(\overline{x}_1, \ldots, \overline{x}_k, E_w, \ldots, E_w, \underline{x}_{k+l}, \ldots, \overline{x}_n)$. So, by considering $2n+2$ zones, we can find $\underline{V}_w$ in time $O(n^2)$.

For $\overline{V}_w$, we can prove that the maximum is always attained at values $x_i = \underline{x}_i$ or $x_i = \overline{x}_i$, so we can always find it in time $O(2^n)$. If no more than K intervals intersect, then, similarly to the non-weighted variance, we can compute $\overline{V}_w$ in time $O(n^2)$.

Robust estimates for the mean. Arithmetic average is vulnerable to outliers: if one of the values is accidentally mis-read as 10^6 times larger than the others, the average is ruined. Several techniques have been proposed to make estimates robust; see, e.g., (Huber, 2004). The best known estimate of this type is the median; there are also more general *L-estimates* of the type $\sum\limits_{i=1}^{n} w_i \cdot x_{(i)}$, where $w_1 \geq 0, \ldots, w_n \geq 0$ are given constants, and $x_{(i)}$ is the i-th value in the ordering of $x_1, \ldots, x_n$ in increasing order. Other techniques include *M-estimates*, i.e., estimates a for which $\sum\limits_{i=1}^{n} \psi(|x_i - a|) \to \max\limits_{a}$ for some non-decreasing function $\psi(x)$.

Each of these statistics C is a (non-strictly) increasing function of each of the variables x_i. Thus, similarly to the average, $\mathbf{C} = [C(\underline{x}_1, \ldots, \underline{x}_n), C(\overline{x}_1, \ldots, \overline{x}_n)]$.

Robust estimates for the generalized central moments. When we discussed central moments, we considered generalized central moments $M_\psi = \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} \psi(x_i - E)$ for an appropriate convex function $\psi(x)$. In that description, we assumed that E is the usual average.

It is also possible to consider the case when E is not the average, but the value for which $\sum\limits_{i=1}^{n} \psi(x_i - E) \to \min\limits_{E}$. In this case, the robust estimate for the generalized central moment takes the form

$$M_\psi^{\text{rob}} = \min_E \left(\frac{1}{n} \cdot \sum_{i=1}^{n} \psi(x_i - E) \right).$$

Since the function $\psi(x)$ is convex, the expression $\sum\limits_{i=1}^{n} \psi(x_i - E)$ is also convex, so it only attains its maximum at the vertices of the convex box $\mathbf{x}_1 \times \ldots \times \mathbf{x}_b$, i.e., when for every i, either $x_i = \underline{x}_i$ or $x_i = \overline{x}_i$. For the case of a single MI, the same proof as for the average E enables us to conclude that the maximum of the new generalized central moment is also always attained at one of n vectors $(\underline{x}_1, \ldots, \underline{x}_k, \overline{x}_{k+1}, \ldots, \overline{x}_n)$,

and thus, that this maximum can be computed in time $O(n^2)$. For m MIs, we need time $O(n^{m+1})$.

Correlation. For correlation, we only know that in general, the problem of computing the exact range is NP-hard (Ferson et al., 2002d).

4. Additional Issues

On-line data processing. In the above text, we implicitly assumed that before we start computing the statistics, we have all the measurement results. In real life, we often continue measurements after we started the computations. Traditional estimates for mean and variance can be easily modified with the arrival of the new measurement result x_{n+1}: $E' = (n \cdot E + x_{n+1})/(n+1)$ and $V' = M' - (E')^2$, where $M' = (n \cdot M + x_{n+1}^2)/(n+1)$ and $M = V + E^2$. For the interval mean, we can have a similar adjustment. However, for other statistics, the above algorithms for processing interval data require that we start computation from scratch. Is it possible to modify these algorithms to adjust them to on-line data processing? The only statistic for which such an adjustment is known is the variance, for which an algorithm proposed in (Wu et al., 2003; Kreinovich et al., (in press)) requires only $O(n)$ steps to incorporate a new interval data point.

In this algorithm, we store the sorting corresponding to the zones and we store auxiliary results corresponding to each zone (finitely many results for each zone). So, if only d out of n intervals are non-degenerate, we only need $O(d)$ steps to incorporate a new data point.

Fuzzy data. Often, in addition to (or instead of) the guaranteed bounds, an expert can provide bounds that contain x_i with a certain degree of confidence. Often, we know several such bounding intervals corresponding to different degrees of confidence. Such a nested family of intervals is also called a *fuzzy set*, because it turns out to be equivalent to a more traditional definition of fuzzy set (Nguyen and Kreinovich, 1996; Nguyen and Walker, 1999) (if a traditional fuzzy set is given, then different intervals from the nested family can be viewed as α-cuts corresponding to different levels of uncertainty α).

To provide statistical analysis of fuzzy-valued data, we can therefore, for each level α, apply the above interval-valued techniques to the corresponding α-cuts (Martinez, 2003; Nguyen et al., 2003).

Can we detect the case of several MI? For the several MI case, we assumed that measurement are labeled, so that we can check which

measurements were done by each MI; this labeling is used in the algorithms. What if we do not keep records on which interval was measured by which MI; can we then reconstruct the labels and thus apply the algorithms?

For two MI, we can: we pick an interval and call it MI_1. If any other interval is in subset relation with this one, then this new interval is MI_2. At any given stage, if one of the un-classified intervals is in subset relation with one of the already classified ones, we classify it to the opposite class. If none of the un-classified intervals is in subset relation with classified ones, we pick one of the un-classified ones and assign to MI_1. After $\leq n$ iterations, we get the desired labeling.

In general, for m MI, the labeling may not be easy. Indeed, we can construct a graph in which vertices are intervals, and vertices are connected if they are in a subset relation. Our objective is to assign a class to each vertex so that connected vertices cannot be of the same class. This is exactly the coloring problem that is known to be NP-hard (Garey and Johnson, 1979).

Parallelization. In the general case, the problem of computing the range $\mathbf{C}$ of a statistic C on interval data $\mathbf{x}_i$ requires too much computation time. One way to speed up computations is to use parallel computations.

If we have a potentially unlimited number of parallel processors, then, for the mean, the addition can be done in time $O(\log(n))$ (Jaja, 1992). In $O(n{\cdot}\log(n))$ and $O(n^2)$ algorithms for computing $\underline{V}$ and $\overline{V}$, we can perform sorting in time $O(\log(n))$, then compute V_k for each zone in parallel, and find the largest of the n resulting values V_k in parallel (in time $O(\log(n))$). The sum that constitutes the variance can also be computed in parallel in time $O(\log(n))$, so overall, we need $O(\log(n))$ time.

Similarly, we can transform polynomial algorithms for computing the bounds for covariance, outlier statistics (L, U, and R), and moments into $O(\log(n))$ parallel algorithms.

In the general case, to find $\overline{V}$ and other difficult-to-compute bounds, we must compute the largest of the $N \stackrel{\mathrm{def}}{=} 2^n$ values corresponding to 2^n possible combinations of $\underline{x}_i$ and $\overline{x}_i$. This maximum can be computed in time $O(\log(N)) = O(n)$. This does not mean, of course, that we can always physically compute $\overline{V}$ in linear time: communication time grows exponentially with n; see, e.g., (Morgenstein and Kreinovich, 1995).

It is desirable to also analyze the case when we have a limited number of processors $p \ll n$.

Quantum algorithms. Another way to speed up computations is to use quantum computing. In (Martinez, 2003; Kreinovich and Longpré, 2004), we describe how quantum algorithms can speed up the computation of $\mathbf{C}$.

What if we have partial information about the probabilities? Enter p-boxes. In the above text, we assumed that the only information that we have about the measurement error Δx is that this error is somewhere in the interval $[-\Delta, \Delta]$, and that we have no information about the probabilities of different values from this interval. In many real-life situations, we do not know the exact probability distribution for Δx, but we have a partial information about the corresponding probabilities. How can we describe this partial information?

To answer this question, let us recall how the complete information about the probability distribution is usually described. A natural way to describe a probability distribution is by describing its cumulative density function (cdf) $F(t) \stackrel{\text{def}}{=} \text{Prob}(\Delta x \le t)$. In practice, a reasonable way to store the information about $F(t)$ is to store *quantiles*, i.e., to fix a natural number n and to store, for every i from 0 to n, the values t_i for which $F(t_i) = i/n$. Here, t_0 is the largest value for which $F(t_0) = 0$ and t_n is the smallest value for which $F(t_n) = 1$, i.e., $[t_0, t_n]$ is the smallest interval on which the probability distribution is located with probability 1.

If we only have partial information about the probabilities, this means that – at least for some values t – we do not know the exact value of $F(t)$. At best, we know an interval $\mathbf{F}(t) = [\underline{F}(t), \overline{F}(t)]$ of possible values of $F(t)$. So, a natural way to describe partial information about the probability distribution is to describe the two functions $\underline{F}(t)$ and $\overline{F}(t)$. This pair of cdfs is called a *p-box*; see, e.g., a book (Ferson, 2002). In addition to the theoretical concepts, this book describes the software tool for processing different types of uncertainty, a tool based on the notion of a p-box.

Similarly to the case of full information, it is reasonable to store the corresponding quantiles, i.e., the values $\underline{t}_i$ for which $\overline{F}(\underline{t}_i) = i/n$ and the values $\overline{t}_i$ for which $\underline{F}(\overline{t}_i) = i/n$. (The reason why we switched the notations is because $\underline{F}(t) \le \overline{F}(t)$ implies $\underline{t}_i \le \overline{t}_i$.) This is exactly the representation used in (Ferson, 2002).

What if we have partial information about the probabilities? Processing p-boxes and how the above alorithms can help. Once we have a probability distribution $F(t)$, natural questions are: what is the mean and the variance of this distribution? A p-box means that several different distributions are possible, and for different distributions, we may have

different values of means and variance. So, when we have a p-box, natural questions are: what is the range of possible values of the mean? what is the range of possible values of the variance?

The mean E is a monotonic function of $F(t)$; so, for the mean E, the answer is simple: the mean of $\underline{F}(t)$ is the desired upper bound $\overline{E}$ for E, and the mean of $\overline{F}(t)$ is the desired lower bound $\underline{E}$ for E. The variance V is not monotonic, so the problem of estimating the variance is more difficult.

For the case of the exact distribution, if we have the quantiles $t(\alpha)$ corresponding to all possible probability values $\alpha \in [0,1]$, then we can describe the mean of the corresponding probability distribution as $E = \int t(\alpha)\,d\alpha$, and the variance as $V = \int (t(\alpha) - E)^2\,d\alpha$. If we only know the quantiles $t_1 = t(1/n), \ldots, t_n = t(n/n)$, then it is reasonable to replace the integral by the corresponding integral sum; as a result, we get the estimates $E = \dfrac{1}{n}\sum\limits_{i=1}^{n} t_i$ and $V = \dfrac{1}{n}\sum\limits_{i=1}^{n} (t_i - E)^2$.

In these terms, a p-box means that instead of the exact value t_i of each quantile, we have an interval of possible values $[\underline{t}_i, \overline{t}_i]$. So, to find the range of V, we must consider the range of possible values of V when $t_i \in [\underline{t}_i, \overline{t}_i]$. There is an additional restriction that the values t_i should be (non-strictly) increasing: $t_i \le t_{i+1}$.

The resulting problem is very similar to the problems of estimating mean and variance of the interval data. In this case, intervals satisfy the subset property, i.e., we are in the case that we called the case of single MI. The only difference between the current problem of analyzing p-boxes and the above problem is that in the above problem, we looked for minimum and maximum of the variance over all possible vectors x_i for which $x_i \in \mathbf{x}_i$ for all i, while in our new problem, we have an additional monotonicity restriction $t_i \le t_{i+1}$. However, the solutions to our previous problems of computing $\underline{V}$ and $\overline{V}$ for the case of a single MI are actually attained at vectors that are monotonic. Thus, to find the desired value V, we can use the same algorithm as we described above.

Specifically, to find $\underline{V}$, we find k for which the variance of the vector $t = (\overline{t}_1, \ldots, \overline{t}_k, \bar{t}, \ldots, \bar{t}, \underline{t}_{k+l}, \ldots, \underline{t}_n)$ for which the variance is the smallest. To find $\overline{V}$, we find k for which the variance of the vector $t = (\underline{t}_1, \ldots, \underline{t}_k, \overline{t}_{k+1}, \ldots, \overline{t}_n)$ for which the variance is the largest. Intuitively, this makes perfect sense: to get the smallest V, we select the values t_i as close to the average $\bar{t}$ as possible; to get the largest V, we select the values t_i as far away from the average $\bar{t}$ as possible. In both case, we can compute $\underline{V}$ and $\overline{V}$ in time $O(n \cdot \log(n))$.

The above algorithm describes a heuristic estimate based on approximating an integral with an integral sum. To get reliable bounds, we can

take into consideration that both bounds $\underline{F}(t)$ and $\overline{F}(t)$ are monotonic; thus, we can always replace the p-box by a larger p-box in which the values $t(\alpha)$ are piecewise-constant: namely, we take $\mathbf{t}_i' = [\underline{t}_{i-1}, \overline{t}_i]$ for each i. For this new p-box, the integral sum coincides with the integral, so the range $[\underline{V}, \overline{V}]$ produced by the above algorithm is exactly the range of the variance over all possible distributions from the enlarged p-box. It is therefore guaranteed to contain the range of possible values of the variance V for the original p-box.

What if we have partial information about probabilities? Multi-dimensional case. How can we describe partial information about probabilities in multi-dimensional case? A traditional analogue of a cdf is a multi-dimensional cdf

$$F(t_1, \ldots, t_p) = \mathrm{Prob}(x_1 \le t_1 \,\&\, \ldots \,\&\, x_p \le t_p);$$

see, e.g., (Wadsworth, 1990). The problem with this definition is that often multi-D data represent, e.g., vectors with components $x_1, \ldots, x_p$. The components depend on the choice of coordinates. As a result, even if a distribution is symmetric – e.g., a rotation-invariant Gaussian distribution – the description in terms of a multi-D cdf is *not* rotation-invariant.

It is desirable to come up with a representation that preserves such a symmetry. A natural way to do it is to store, for each half-space, the probability that the vector $\vec{x} = (x_1, \ldots, x_p)$ is within this half-space. In other words, for every unit vector $\vec{e}$ and for every value t, we store the probability $F(\vec{e}, t) \stackrel{\mathrm{def}}{=} \mathrm{Prob}(\vec{x} \cdot \vec{e} \le t)$, where $\vec{x} \cdot \vec{e} = x_1 \cdot e_1 + \ldots + x_n \cdot e_n$ is a scalar (dot) product of the two vectors. This representation is clearly rotation-invariant: if we change the coordinates, we keep the same values $F(\vec{e}, t)$; the only difference is that we store each value under different (rotated) $\vec{e}$. Moreover, this representation is invariant under arbitrary linear transformations.

Based on this information, we can uniquely determine the probability distribution. For example, if the probability distribution has a probability density function (pdf) $\rho(\vec{x})$, then this pdf can be reconstructed as follows. First, we determine the characteristic function $\chi(\vec{\omega}) \stackrel{\mathrm{def}}{=} E[\exp(\mathrm{i} \cdot (\vec{x} \cdot \vec{\omega}))]$, where $E[\cdot]$ stands for the expected value. To get the value of $\chi(\vec{\omega})$, we apply the 1-D Fourier transform, to the values $F(\vec{e}, t)$ for different t, where $\vec{e} \stackrel{\mathrm{def}}{=} \vec{\omega}/\|\vec{\omega}\|$ is a unit vector in the direction of $\vec{\omega}$. Then, we can find $\rho(\vec{x})$ by applying the p-dimensional Inverse Fourier Transform to $\chi(\vec{\omega})$.

It is therefore reasonable to represent a partial information about the probability distribution by storing, for each $\vec{e}$ and t, the bounds $\underline{F}(\vec{e},t)$ and $\overline{F}(\vec{e},t)$ that describe the range of possible values for $F(\vec{e},t)$.

It is worth mentioning that since for continuous distributions, $F(\vec{e},t) = 1 - F(-\vec{e},-t)$, we have $\underline{F}(\vec{e},t) = 1 - \overline{F}(-\vec{e},-t)$. So, it is sufficient to only describe $\overline{F}(\vec{e},t)$, the lower bounds $\underline{F}(\vec{e},t)$ can then be uniquely determined (or, vice versa, we can only describe the values $\underline{F}(\vec{e},t)$; then the values $\overline{F}(\vec{e},t)$ will be uniquely determined).

In order to transform this idea into an efficient software tool, we need to solve two types of problems. First, we must solve algorithmic problems: develop algorithms for estimating the ranges of statistical characteristics (such as moments) for the corresponding multi-D p-boxes.

Second, we must solve implementation problems. Theoretically, to uniquely describe a probability distribution, we need to know infinitely many values $F(\vec{e},t)$ corresponding to infinitely many different vectors $\vec{e}$ and infinitely many different numbers t. In practice, we can only store finitely many values $F(\vec{e},t)$ corresponding to finitely many vectors $\vec{e}$.

In principle, we can simply select a rectangular grid and store the values for the vectors $\vec{e}$ from this grid. However, the selection of the grid violates rotation-invariance and thus, eliminates the advantage of selecting this particular multi-D analogue of a cdf. It turns out that there is a better way: instead of using a grid, we can use rational points on a unit sphere. There exists efficient algorithms for generating such points, and the set of all such points is almost rotation-invariant: it is invariant with respect to all rotations for which all the entries in the corresponding rotation matrix are rational numbers (Oliverio, 1996; Trautman, 1998).

Beyond p-boxes? A p-box does not fully describe all kinds of possible partial information about the probability distribution. For example, the same p-box corresponds to the class of all distributions located on an interval $[0,1]$ and to the class of all distributions located at two points 0 and 1.

Similarly, in the multi-D case, if we only use the above-defined multi-D cdfs, we will not be able to distinguish between a set S (= the class of all probability distributions localized on the set S with probability 1) and its convex hull. To provide such a distinction, we may want, in addition to the bounds on the probabilities $\mathrm{Prob}(f(x) \le t)$ for all *linear* functions $f(x)$, to also keep the bounds on the similar probabilities corresponding to all *quadratic* functions $f(x)$.

Let us show that this addition indeed enables us to distinguish between different sets S. Indeed, for every point x, to check whether

$x \in S$, we ask, for different values $\varepsilon > 0$, for the upper bound for the probability $\mathrm{Prob}(d^2(x, x_0) \leq \varepsilon^2)$, where $d(x, x_0)$ denotes the distance between the two points. If $x \notin S$, then for sufficiently small ε, this probability will be 0; on the other hand, if $x \in S$, then it is possible that we have a distribution located at this point x with probability 1, so the upper bound is 1 for all ε (Nguyen et al., 2000).

In 1-D case, the condition $f(x) \leq t$ for a non-linear quadratic function $f(x)$ is satisfied either inside an interval, or outside an interval. Thus, in 1-D case, our idea means that in addition to cdf, we also store the bounds on the probabilities of x being within different intervals. Such bounds are analyzed, e.g., in (Berleant, 1993; Berleant, 1996; Berleant and Goodman-Strauss, 1998; Berleant et al., 2003).

Acknowledgements

This work was supported in part by NASA under cooperative agreement NCC5-209, by the Future Aerospace Science and Technology Program (FAST) Center for Structural Integrity of Aerospace Systems, effort sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant F49620-00-1-0365, by NSF grants EAR-0112968, EAR-0225670, and EIA-0321328, by the Army Research Laboratories grant DATM-05-02-C-0046, and by a research grant from Sandia National Laboratories as part of the Department of Energy Accelerated Strategic Computing Initiative (ASCI).

The authors are greatly thankful to the anonymous referees for helpful suggestions.

References

Beck, B., V. Kreinovich, and B. Wu, Interval-Valued and Fuzzy-Valued Random Variables: From Computing Sample Variances to Computing Sample Covariances, In: M. Lopez, M. A. Gil, P. Grzegorzewski, O. Hrynewicz, and J. Lawry, editor, *Soft Methodology and Random Information Systems*, pages 85–92, Springer-Verlag, Berlin Heidelberg New York Tokyo, 2004.

Berleant, D., Automatically verified arithmetic with both intervals and probability density functions, *Interval Computations*, 1993, (2):48–70.

Berleant, D., Automatically verified arithmetic on probability distributions and intervals, In: R. B. Kearfott and V. Kreinovich, editors, *Applications of Interval Computations*, Kluwer, Dordrecht, 1996.

Berleant, D., and C. Goodman-Strauss, Bounding the results of arithmetic operations on random variables of unknown dependency using intervals, *Reliable Computing*, 1998, 4(2):147–165.

Berleant, D., L. Xie, and J. Zhang, Statool: A Tool for Distribution Envelope Determination (DEnv), an Interval-Based Algorithm for Arithmetic on Random Variables, *Reliable Computing*, 2003, 9(2):91–108.

Ferson, S. *RAMAS Risk Calc 4.0: Risk Assessment with Uncertain Numbers*, CRC Press, Boca Raton, Florida, 2002.

Ferson, S., L. Ginzburg, V. Kreinovich, and M. Aviles, Exact Bounds on Sample Variance of Interval Data. *Extended Abstracts of the 2002 SIAM Workshop on Validated Computing*, Toronto, Canada, 2002, pp. 67–69

Ferson, S., L. Ginzburg, V. Kreinovich, and M. Aviles, *Exact Bounds on Finite Populations of Interval Data*, University of Texas at El Paso, Department of Computer Science, Technical Report UTEP-CS-02-13d, 2002, http://www.cs.utep.edu/vladik/2002/tr02-13d.pdf

Ferson, S., L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, Computing Variance for Interval Data is NP-Hard, *ACM SIGACT News*, 33(2):108–118, 2002.

Ferson, S., D. Myers, and D. Berleant, *Distribution-free risk analysis: I. Range, mean, and variance*, Applied Biomathematics, Technical Report, 2001.

Garey, M. E., and D. S. Johnson, *Computers and intractability: a guide to the theory of NP-completeness*, Freeman, San Francisco, 1979.

Granvilliers, L., V. Kreinovich, and N. Müller, Novel Approaches to Numerical Software with Result Verification", In: R. Alt, A. Frommer, R. B. Kearfott, and W. Luther, editors, *Numerical Software with Result Verification*, International Dagstuhl Seminar, Dagstuhl Castle, Germany, January 19–24, 2003, Revised Papers, Springer Lectures Notes in Computer Science, 2004, Vol. 2991, pp. 274–305.

Huber, P. J., *Robust statistics*, Wiley, New York, 2004.

Jájá, J. *An Introduction to Parallel Algorithms*, Addison-Wesley, Reading, MA, 1992.

Kreinovich, V. Probabilities, Intervals, What Next? Optimization Problems Related to Extension of Interval Computations to Situations with Partial Information about Probabilities, *Journal of Global Optimization* (in press).

Kreinovich, V., A. Lakeyev, J. Rohn, and P. Kahl, *Computational complexity and feasibility of data processing and interval computations*, Kluwer, Dordrecht, 1997.

Kreinovich, V., and L. Longpré, Computational complexity and feasibility of data processing and interval computations, with extension to cases when we have partial information about probabilities, In: V. Brattka, M. Schroeder, K. Weihrauch, and N. Zhong, editors, *Proc. Conf. on Computability and Complexity in Analysis CCA'2003*, Cincinnati, Ohio, USA, August 28–30, 2003, pp. 19–54.

Kreinovich, V., and L. Longpré, Fast Quantum Algorithms for Handling Probabilistic and Interval Uncertainty, *Mathematical Logic Quarterly*, 2004, 50(4/5):507–518.

Kreinovich, V., L. Longpré, S. Ferson, and L. Ginzburg, *Computing Higher Central Moments for Interval Data*, University of Texas at El Paso, Department of Computer Science, Technical Report UTEP-CS-03-14b, 2004, http://www.cs.utep.edu/vladik/2003/tr03-14b.pdf

Kreinovich, V., L. Longpré, P. Patangay, S. Ferson, and L. Ginzburg, Outlier Detection Under Interval Uncertainty: Algorithmic Solvability and Computational Complexity, In: I. Lirkov, S. Margenov, J. Wasniewski, and P. Yalamov, editors, *Large-Scale Scientific Computing*, Proceedings of the 4-th International Conference LSSC'2003, Sozopol, Bulgaria, June 4–8, 2003, Springer Lecture Notes in Computer Science, 2004, Vol. 2907, pp. 238–245

Kreinovich, V., H. T. Nguyen, and B. Wu, On-Line Algorithms for Computing Mean and Variance of Interval Data, and Their Use in Intelligent Systems, *Information Sciences* (in press).

Kreinovich, V., P. Patangay, L. Longpré, S. A. Starks, C. Campos, S. Ferson, and L. Ginzburg, Outlier Detection Under Interval and Fuzzy Uncertainty: Algorithmic Solvability and Computational Complexity, *Proceedings of the 22nd International Conference of the North American Fuzzy Information Processing Society NAFIPS'2003*, Chicago, Illinois, July 24–26, 2003, pp. 401–406.

Kuznetsov, V. P., *Interval Statistical Models*, Radio i Svyaz, Moscow, 1991 (in Russian).

Lodwick, W. A., and K. D. Jamison, Estimating and Validating the Cumulative Distribution of a Function of Random Variables: Toward the Development of Distribution Arithmetic, *Reliable Computing*, 2003, 9(2):127–141.

Martinez, M., L. Longpré, V. Kreinovich, S. A. Starks, and H. T. Nguyen, Fast Quantum Algorithms for Handling Probabilistic, Interval, and Fuzzy Uncertainty, *Proceedings of the 22nd International Conference of the North American Fuzzy Information Processing Society NAFIPS'2003*, Chicago, Illinois, July 24–26, 2003, pp. 395–400.

Moore, R. E., and W. A. Lodwick, Interval Analysis and Fuzzy Set Theory, *Fuzzy Sets and Systems*, 2003, 135(1):5–9.

Morgenstein, D., and V. Kreinovich, Which algorithms are feasible and which are not depends on the geometry of space-time, *Geombinatorics*, 1995, 4(3):80–97.

Nguyen, H. T., and V. Kreinovich, Nested Intervals and Sets: Concepts, Relations to Fuzzy Sets, and Applications, In: R. B. Kearfott and V. Kreinovich, editors, *Applications of Interval Computations*, Kluwer, Dordrecht, 1996, pp. 245–290

Nguyen, H. T., and E. A. Walker, *First Course in Fuzzy Logic*, CRC Press, Boca Raton, Florida, 1999.

Nguyen, H. T., T. Wang, and V. Kreinovich, Towards Foundations of Processing Imprecise Data: From Traditional Statistical Techniques of Processing Crisp Data to Statistical Processing of Fuzzy Data, In: Y. Liu, G. Chen, M. Ying, and K.-Y. Cai, editors, *Proceedings of the International Conference on Fuzzy Information Processing: Theories and Applications FIP'2003*, Beijing, China, March 1–4, 2003, Vol. II, pp. 895–900.

Nguyen, H. T., B. Wu, and V. Kreinovich, Shadows of Fuzzy Sets – A Natural Approach Towards Describing 2-D and Multi-D Fuzzy Uncertainty in Linguistic Terms, *Proc. 9th IEEE Int'l Conference on Fuzzy Systems FUZZ-IEEE'2000*, San Antonio, Texas, May 7–10, 2000, Vol. 1, pp. 340–345.

Nivlet, P., F. Fournier, and J. Royer, A new methodology to account for uncertainties in 4-D seismic interpretation, *Proc. 71st Annual Int'l Meeting of Soc. of Exploratory Geophysics SEG'2001*, San Antonio, TX, September 9–14, 2001, 1644–1647.

Nivlet, P., F. Fournier, and J. Royer, Propagating interval uncertainties in supervised pattern recognition for reservoir characterization, *Proc. 2001 Society of Petroleum Engineers Annual Conf. SPE'2001*, New Orleans, LA, September 30–October 3, 2001, paper SPE-71327.

Oliverio, P., Self-generating Pythagorean quadruples and n-tuples, *Fibonacci Quarterly*, May 1996, pp. 98–101.

Osegueda, R., V. Kreinovich, L. Potluri, R. Aló, Non-Destructive Testing of Aerospace Structures: Granularity and Data Mining Approach, *Proc. FUZZ-IEEE'2002*, Honolulu, HI, May 12–17, 2002, Vol. 1, pp. 685–689

Rabinovich, S., *Measurement Errors: Theory and Practice*, American Institute of Physics, New York, 1993.

Regan, H., S. Ferson, and D. Berleant, Equivalence of five methods for bounding uncertainty, *Journal of Approximate Reasoning* (in press).

Rowe, N. C., Absolute bounds on the mean and standard deviation of transformed data for constant-sign-derivative transformations, *SIAM Journal of Scientific Statistical Computing*, 1988, 9:1098–1113.

Shmulevich, I., and W. Zhang, Binary analysis and optimization-based normalization of gene expression data, *Bioinformatics*, 2002, 18(4):555–565.

Trautman, A., Pythagorean Spinors and Penrose Twistors, In: S. A. Hugget et al., editors, *The Geometric Universe; Science, Geometry, and the Work of Roger Penrose*, Oxford Univ. Press, Oxford, 1998.

Wadsworth, H. M. Jr., editor, *Handbook of statistical methods for engineers and scientists*, McGraw-Hill Publishing Co., N.Y., 1990.

Wu, B., H. T. Nguyen, and V. Kreinovich, Real-Time Algorithms for Statistical Analysis of Interval Data, *Proceedings of the International Conference on Information Technology InTech'03*, Chiang Mai, Thailand, December 17–19, 2003, pp. 483–490.

Walley, P., *Statistical Reasoning with Imprecise Probabilities*, Chapman & Hall, N.Y., 1991.

Williamson, R., and T. Downs, Probabilistic arithmetic I: numerical methods for calculating convolutions and dependency bounds, *International Journal of Approximate Reasoning*, 1990, 4:89–158.

Xiang, G., S. A. Starks, V. Kreinovich, and L. Longpré, New Algorithms for Statistical Analysis of Interval Data, *Proceedings of the Workshop on State-of-the-Art in Scientific Computing PARA'04*, Lyngby, Denmark, June 20–23, 2004, Vol. 1, pp. 123–129.

Zhang, W., I. Shmulevich, and J. Astola, *Microarray Quality Control*, Wiley, Hoboken, New Jersey, 2004.