12-1-2006

Fast Algorithms for Computing Statistics under Interval and Fuzzy Uncertainty, and Their Applications

Gang Xiang

Vladik Kreinovich
University of Texas at El Paso, vladik@utep.edu

Comments:
UTEP-CS-06-48.
Published in *Proceedings of the International Conference on Fuzzy Mathematics and Its Applications*, Ahmednagar, Maharashtra, India, January 27-29, 2007.

Fast Algorithms for Computing Statistics Under Interval and Fuzzy Uncertainty, and Their Applications

Gang Xiang and Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso
El Paso, TX 79968, USA, {gxiang,vladik}@utep.edu

Abstract

In many engineering applications, we have to combine probabilistic, interval, and fuzzy uncertainty. 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 measure the values with interval uncertainty. We must therefore modify the existing statistical algorithms to process such interval data.

In this paper, we provide a brief survey of algorithms for computing various statistics under interval (and fuzzy) uncertainty and of their applications, including applications to the seismic inverse problem in geosciences, to chip design in computer engineering, and to radar data processing.

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 $E(x_1,\ldots,x_n)=\dfrac{1}{n}\cdot(x_1+\ldots+x_n)$; a reasonable statistic for estimating the variance V is the population variance $V(x_1,\ldots,x_n)=\dfrac{1}{n}\cdot\sum\limits_{i=1}^{n}(x_i-E)^2$.

Interval Uncertainty. In environmental measurements, we often only measure 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 distribution of different values in 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., [33, 34]). 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 $\Delta x\stackrel{\rm def}{=}\widetilde{x}-x$ (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 true 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 [29, 30]. Detecting outliers is also important in bioinformatics [37].

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., [43].

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., [20, 42].

Fuzzy Uncertainty. In addition to measurement results, we often have expert estimates of the desired quantities. An expert often describes such an estimate by using words from the natural language, like "most probably, the value of the quantity is between 6 and 7, but it is somewhat possible to have values between 5 and 8". To formalize this knowledge, it is natural to use *fuzzy set theory*, a formalism specifically designed for describing this type of informal ("fuzzy") knowledge [18, 27].

As a result, for every value x_i, we have a fuzzy set $\mu_i(x_i)$ which describes the expert's prior knowledge about x_i: the number $\mu_i(x_i)$ describes the expert's degree of certainty that x_i is a possible value of the i-th quantity.

An alternative user-friendly way to represent a fuzzy set is by using its α-cuts $\{x_i \,|\, \mu_i(x_i) > \alpha\}$ (or $\{x_i \,|\, \mu_i(x_i) \geq \alpha\}$). For example, the α-cut corresponding to $\alpha = 0$ is the set of all the values which are possible at all, the α-cut corresponding to $\alpha = 0.1$ is the set of all the values which are possible with degree of certainty at least 0.1, etc. In these terms, a fuzzy set can be viewed as a nested family of intervals $[\underline{x}_i(\alpha), \overline{x}_i(\alpha)]$ corresponding to different level α.

Estimating Statistics under Fuzzy Uncertainty: Precise Formulation of the Problem. In general, we have fuzzy knowledge $\mu_i(x_i)$ about each value x_i; we want to find the fuzzy set corresponding to a given characteristic $y = C(x_1, \ldots, x_n)$. Intuitively, the value y is a reasonable value of the characteristic if $y = f(x_1, \ldots, x_n)$ for some reasonable values x_i, i.e., if for some values $x_1, \ldots, x_n$, x_1 is reasonable, and x_2 is reasonable, ..., and $f = f(x_1 \ldots, x_n)$. If we interpret "and" as min and "for some" ("or") as max, then we conclude that the corresponding degree of certainty $\mu(y)$ in y is equal to $\mu(y) = \max\{\min(\mu_1(x_1), \ldots, \mu_n(x_n)) | C(x_1, \ldots, x_n) = y\}$.

Reduction to the Case of Interval Uncertainty. It is know that the above formula (called *extension principle*) can be reformulated as follows: for each α, the α-cut $\mathbf{y}(\alpha)$ of y is equal to the range of possible values of $C(x_1, \ldots, x_n)$ when $x_i \in \mathbf{x}_i(\alpha)$ for all i. Thus, from the computational viewpoint, the problem of computing the statistical characteristic under fuzzy uncertainty can be reduced to the problem of computing this characteristic under interval uncertainty.

Estimating Statistics Under Interval Uncertainty: a Problem. In the case of interval uncertainty, instead of the true 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) true 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 compute, or bound these ranges. This is the problem that we will be analyzing 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 [23, 38]; for further developments, see, e.g., [4, 5, 6, 7, 14, 15, 24, 26, 35, 36, 39] 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, $\underline{E} = \dfrac{1}{n} \cdot (\underline{x}_1 + \ldots + \underline{x}_n)$ and $\overline{E} = \dfrac{1}{n} \cdot (\overline{x}_1 + \ldots + \overline{x}_n)$.

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 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., [19].)

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 $C_{\text{lin}}(x_1, \ldots, x_n) = C_0 - \sum\limits_{i=1}^{n} C_i \cdot \Delta x_i$, 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 $\Delta \stackrel{\text{def}}{=} \sum\limits_{i=1}^{n} |C_i| \cdot \Delta_i$.

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.

For this Problem, Traditional Interval Methods Sometimes Lead to Excess Width. Let us show that for this problem, traditional interval methods sometimes lead to excess width.

Historically the first method for computing the enclosure for the range is the method which is sometimes called “straightforward” interval computations. This method is based on the fact that inside the computer, every algorithm consists of elementary operations (arithmetic operations, min, max, etc.). For each elementary operation $f(a, b)$, if we know the intervals $\mathbf{a}$ and $\mathbf{b}$ for a and b, we can compute the exact range $f(\mathbf{a}, \mathbf{b})$. The corresponding formulas form the so-called *interval arithmetic*. In straightforward interval computations, we repeat the computations forming the program f step-by-step, replacing each operation with real numbers by the corresponding operation of interval arithmetic. It is known that, as a result, we get an enclosure for the desired range.

For the problem of computing the range of finite population average, as we have mentioned, straightforward interval computations lead to exact bounds. The reason: in the above formula for E, each interval variable only occurs once [17].

For the problem of computing the range of finite population variance, the situation is somewhat more difficult, because in the expression for the variance, each variable x_i occurs several times: explicitly, in $(x_i - E)^2$, and explicitly, in the expression for E. In this cases, often, dependence between intermediate computation results leads to excess width of the results of straightforward interval computations. Not surprisingly, we do get excess width when applying straightforward interval computations to the above formula.

A better range is often provided by a *centered form*, in which a range $f(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of a smooth function on a box $\mathbf{x}_1 \times \ldots \times \mathbf{x}_n$ is estimated as $f(\mathbf{x}_1, \ldots, \mathbf{x}_n) \subseteq f(\widetilde{x}_1, \ldots, \widetilde{x}_n) + \sum\limits_{i=1}^{n} \dfrac{\partial f}{\partial x_i}(\mathbf{x}_1, \ldots, \mathbf{x}_n) \cdot [-\Delta_i, \Delta_i]$, where $\widetilde{x}_i = (\underline{x}_i + \overline{x}_i)/2$ is the interval's midpoint and $\Delta_i = (\underline{x}_i - \overline{x}_i)/2$ is its half-width. However, this formula also leads to excess width.

For this Problem, Traditional Optimization Methods Sometimes Require Unreasonably Long Time. A natural way to solve the problem of computing the exact range $[\underline{V}, \overline{V}]$ of the finite population variance is to solve it as a constrained optimization problem. Specifically, to find $\underline{V}$, we must find the minimum of the function (1.1) under the conditions $\underline{x}_1 \le x_1 \le \overline{x}_1$, ..., $\underline{x}_n \le x_n \le \overline{x}_n$. Similarly, to find $\overline{V}$, we must find the maximum of the function (1.1) under the same conditions.

There exist optimization techniques that lead to computing "sharp" (exact) values of $\min(f(x))$ and $\max(f(x))$. However, the behavior of such general constrained optimization algorithms is not easily predictable, and can, in general, be exponential in n.

For small n, this is quite doable, but for large n, the exponential computation time grows so fast that for reasonable n, it becomes unrealistically large: e.g., for $n \approx 300$, it becomes larger than the lifetime of the Universe.

We Need New Methods. Summarizing: the existing methods are either not always efficient, or do not always provide us with sharp estimates for $\underline{V}$ and $\overline{V}$. So, we need new methods.

3 Reasonable Classes of Problems for Which We Can Expect Feasible Algorithms for Statistics of Interval Data

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

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.

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).

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: $[\underline{x}_i, \overline{x}_i] \subseteq (\underline{x}_j, \overline{x}_j)$. We therefore say that a collection of intervals satisfies a *subset property* if $[\underline{x}_i, \overline{x}_i] \not\subseteq (\underline{x}_j, \overline{x}_j)$ for all i and j for which the intervals $\mathbf{x}_i$ and $\mathbf{x}_j$ are non-degenerate.

Several MI. A 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.

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.

We will call interval data with this property – that every two non-degenerate intervals either coincide or intersect at most in one point – *privacy case.* Such intervals satisfy the subset property.

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.

4 Algorithms

Let us now describe algorithms for computing statistical characteristics under interval uncertainty [10, 11, 21, 22, 40].

Variance: Lower Bound. The lower bound $\underline{V}$ can be always computed in time $O(n \cdot \log(n))$.

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 the minimum is attained at the right endpoint $x_i = \overline{x}_i$ and $\dfrac{\partial V}{\partial x_i} < 0$. Since the partial derivative is equal to $(2/n) \cdot (x_i - E)$, we conclude that either $x_i = E$, or $x_i = \underline{x}_i > E$, or $x_i = \underline{x}_i < E$. Thus, if we know where E is located in relation to all the endpoints, we can uniquely determine the corresponding minimizing value x_i for every i: if $\overline{x}_i \le E$ then $x_i = \overline{x}_i$; if $x_i \le \underline{x}_i$, then $x_i = \underline{x}_i$; otherwise, $x_i = E$. The corresponding value E can be found from the condition that E 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.

As we have mentioned, the corresponding value E can be found from the condition that E is the average of all the selected values x_i. If E is in the zone $[x_{(k)}, x_{(k+1)}]$, then we know all the values x_i, so $n \cdot E$ should be equal to the sum of these values: $n \cdot E = \sum\limits_{i:\underline{x}_i \ge x_{(k+1)}} \underline{x}_i + (n - N_k) \cdot E + \sum\limits_{j:\overline{x}_j \le x_{(k)}} \overline{x}_j$, where by N_k, we denoted the total number of such i's for which $\underline{x}_i \ge x_{(k+1)}$ and j's for which $\underline{x}_j \le x_{(k)}$.

Subtracting $(n - N_k) \cdot E$ from both sides of this equality, we conclude that $N_k \cdot E = S_k$, where $S_k \stackrel{\text{def}}{=} \sum\limits_{i:\underline{x}_i \ge x_{(k+1)}} \underline{x}_i + \sum\limits_{j:\overline{x}_j \le x_{(k)}} \overline{x}_j$. If $N_k = 0$, this means that $x_i = E$ for all i, so $V = 0$. If $N_k \ne 0$, then $E = S_k / N_k$.

Once E is computed, we can now compute the corresponding variance V_k as $M_k - E^2$, where M_k is the second population moment: $M_k = \dfrac{1}{n} \cdot \sum\limits_{i:\underline{x}_i \ge x_{(k+1)}} (\underline{x}_i)^2 + \dfrac{n - N_k}{n} \cdot E^2 + \dfrac{1}{n} \cdot \sum\limits_{j:\overline{x}_j \le x_{(k)}} (\overline{x}_j)^2$, i.e., $V_k = M'_k - \dfrac{N_k}{n} \cdot E^2$, where $M'_k \stackrel{\text{def}}{=} \dfrac{1}{n} \cdot \left(\sum\limits_{i:\underline{x}_i \ge x_{(k+1)}} (\underline{x}_i)^2 + \sum\limits_{j:\overline{x}_j \le x_{(k)}} (\overline{x}_j)^2 \right)$.

Sorting requires $O(n \cdot \log(n))$ steps [9]. Computing the initial values of S_k, N_k, and M'_k requires linear time, i.e., $O(n)$ steps.

For each k, the values S_k, N_k, and M'_k differ from the previous value by only one or two terms – namely, e.g., the values i for which $\underline{x}_i \geq x_{(k)}$ but $\underline{x}_i < x_{(k+1)}$. In other words, the only change is for i for which $x_{(k)} \leq \underline{x}_i < x_{(k+1)}$. Since $x_{(k)}$ is the ordering of all lower and upper bounds, this means that $x_{(k)} = \underline{x}_i$.

Similarly, the only change in the second sum is the term for which $\overline{x}_j = x_{(k)}$.

So, each of these values $S_k, \ldots,$ can be computed from the previous values $S_{k-1}, \ldots$ in a constant number of steps. Thus, the overall number of steps for computing them is linear in n. The smallest of the values V_k is the desired $\underline{V}$. Thus, we can compute $\underline{V}$ in $O(n \cdot \log(n)) + O(n) = O(n \cdot \log(n))$ steps.

In [41], we developed a faster $O(n)$ algorithm. Its main idea is using a linear-time algorithm for computing the median [9] instead of an $O(n \cdot \log(n))$ algorithm for sorting the list.

Variance: Upper Bound. We have already mentioned that computing $\overline{V}$ is, in general, an NP-hard problem. It is known that the maximum of a convex quadratic function on an interval is always attained at one of the endpoints. Thus, $\overline{V}$ is attained when $x_i = \underline{x}_i$ or $x_i = \overline{x}_i$ for each i.

In the case of subset property, 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 $\overline{x}_i \leq \overline{x}_j$). If we replace two values in such a way that their sum (and hence, the overall average) remain unchanged, then the maximum $\overline{V}$ cannot increase. We can thus show that the maximum is attained at one of the vectors $x^{(k)} = (\underline{x}_1, \ldots, \underline{x}_k, \overline{x}_{k+1}, \ldots, \overline{x}_n)$. When we go from a vector $x^{(k)}$ to the vector $x^{(k+1)}$, only one term changes in the vector x, so only one term changes in each of the sums E and M; thus, we need $O(n \cdot \log(n))$ time for sorting and $O(n)$ time for computing all $V(x^{(k)})$ – the total of $O(n \cdot \log(n))$. A linear time algorithm is presented in [41].

4.1 Case of Several MI

In case of several MI, it can be similarly proven that if we sort the intervals corresponding to each MI in lexicographic order, then the maximum of V is attained when from intervals corresponding to each MI, the values x_i corresponding to this MI form a sequence $(\underline{x}_1, \ldots, \underline{x}_{k_j}, \overline{x}_{k_j+1}, \ldots, \overline{x}_{n_j})$, where n_j is the total number of intervals corresponding to the j-th MI.

Thus, to find the maximum of V, we must find the values $k_1, \ldots, k_m$ corresponding to m MIs. For these values, $V = M - E^2$, where $M = \sum M_j$ and $E = \sum E_j$, where we denoted by E_j and M_j, the averages of, correspondingly, x_i and x_i^2, taken by using only results of j-th MI.

For each MI j, we can compute all $n_j + 1$ possible values E_j and M_j in linear time.

There are $\leq n^m$ combinations of k_is; for each combination, we need m additions to compute $E = \sum E_j$, m additions to compute $M = \sum M_j$, and a constant number of operations to compute $V = M - E^2$. Thus, overall, we need time $O(n^m)$.

Cases of Privacy and Non-Detects. Since these two cases are a particular case of the subset property case, any algorithm developed for the subset property case can be applied to these two cases as well.

Other Statistical Characteristics. Similar algorithms are known for covariance, moments, and combinations $E \pm k_0 \cdot \sqrt{V}$ which are important for detecting outliers.

5 Applications

Seismic Inverse Problem. Our civilization greatly depends on the things we extract from the Earth, such as fossil fuels (oil, coal, natural gas), minerals, and water. To be able to distinguish between more promising and less promising locations, it is desirable to determine the structure of the Earth at these locations. To be more precise, we want to know the structure at different depths z at different locations (x, y). In general, to determine the Earth structure, we can use different measurement results that can be obtained without actually drilling the boreholes: e.g., gravity and magnetic measurements, analyzing the travel-times and paths of seismic ways as they propagate through the earth, etc.

To get a better understanding of the Earth structure, we must rely on *active* seismic data – in other words, we must make artificial explosions, place sensors around them, and measure how the resulting seismic waves propagate. The most important information about the seismic wave is the *travel-time* t_i, i.e., the time that it takes for the wave to travel from its source to the sensor. to determine the geophysical structure of a region, we measure seismic travel times and reconstruct velocities at different depths from these data. The problem of reconstructing this structure is called the *seismic inverse problem.*

Once we know the velocities v_j in each cell j, we can then determine the paths which seismic waves take. Seismic waves travel along the shortest path – shortest in terms of time. It can be easily determined that for such paths, within each cell, the path is a straight line, and on the border between the two cells with velocities v and v', the direction of the path changes in accordance with Snell's law $\dfrac{\sin(\varphi)}{v}=\dfrac{\sin(\varphi')}{v'}$, where φ and φ' are the angles between the paths and the line orthogonal to the border between the cells. (If this formula requires $\sin(\varphi')>1$, this means that this wave cannot penetrate into the neighboring cell at all; instead, it bounces back into the original cell with the same angle φ.)

In particular, we can thus determine the paths from the source to each sensor. The travel-time t_i along i-th path can then be determined as the sum of travel-times in different cells j through which this path passes: $t_i=\sum\limits_j \dfrac{\ell_{ij}}{v_j}$, where ℓ_{ij} denotes the length of the part of i-th path within cell j.

There are several algorithms for solving this system of non-linear equations; the most widely used is the following iterative algorithm proposed by John Hole [16].

However, often, the velocity model that is returned by the existing algorithm is not geophysically meaningful: e.g., it predicts velocities outside of the range of reasonable velocities at this depth. To avoid such situations, it is desirable to incorporate the expert knowledge into the algorithm for solving the inverse problem.

There are two types of such knowledge. First, for each cell j, a geophysicist often provides us with his or her estimate of possible values of the corresponding velocity v_j. Sometimes, this estimates comes in the form of an interval $[\underline{v}_j,\overline{v}_j]$ that is guaranteed to contain the (unknown) actual value of slowness. In other cases, it comes in terms of a fuzzy estimate. Such fuzzy prior information can lead to more accurate velocity models; see, e.g., [2, 3, 13, 28].

Second, prior information often comes from processing previous observations of the region of interest. In this case, before our experiments, for each cell j, we know a prior (approximate) velocity value $\widetilde{v}_j$, and we know the accuracy (standard deviation) σ_j of this approximate value $\widetilde{v}_j$. It is known that this prior information can lead to much more accurate velocity models; see, e.g., [25].

In [2], we show how we can modify Hole's algorithm to take into account both interval (fuzzy) and probabilistic uncertainty – e.g., to make sure that the velocities v_j stay within the corresponding intervals.

Chip Design. In chip design, one of the main objectives is to decrease its clock cycle. On the design stage, this time is usually estimated by using worst-case (interval) techniques, in which we only use the bounds on the parameters that lead to delays. This analysis does not take into account that the probability of the worst-case values is usually very small; thus, the resulting estimates are over-conservative, leading to unnecessary over-design and under-performance of circuits. If we knew the *exact* probability distributions of the corresponding parameters, then we could use Monte-Carlo simulations (or the corresponding analytical techniques) to get the desired estimates. In practice, however, we only have *partial* information about the corresponding distributions, and we want to produce estimates that are valid for all distributions which are consistent with this information.

In [31, 32], we describe a general technique that allows us, in particular, to provide such estimates for the clock time.

Radar Data Processing [1, 12] A radar observes the result of a space explosion. Due to radar's low horizontal resolution, we get a 1-D signal $x(t)$ representing different 2-D slices. Based on these slices, we must distinguish between the body at the core of the explosion and the slowly out-moving fragments.

In [12], we describe new algorithms for processing this 1-D data. Since these algorithms are time-consuming, we also exploit the possibility of parallelizing these algorithms.

The shape of the radar signal can provide us with the additional information about the reflecting surface. However, to decrease the noise, radars use filtering, and filtering changes the shapes of the radar signal. It is therefore necessary to reconstruct the original shape of the radar signal. Corresponding algorithms are described in [1].

Halftoning in Image Processing. Inside the computer, a gray-scale image is represented by assigning, to every pixel (n_1, n_2), the intensity $f(n_1, n_2)$ of the color at this pixel. For color images, we must represent the intensity of each color component.

A laser printer cannot print the points of different intensity; at any pixel, it either prints a black (or a colored) dot, or it does not print anything at all. Therefore, when we print an image, we must first transform it into the form $b(n_1, n_2)$ in which at every pixel (n_1, n_2), we only have 0 or 1: 0 if we do not print a black dot at this location, and 1 if we do. This transformation from the original continuous image to the two-level ("halftone") image is called *halftoning*.

Crudely speaking, the level of intensity at a pixel is represented by the relative frequency of black spots around it:

Once we have a printed image, we can digitally scan it and get the halftone values $b(n_1, n_2)$ from this printed page. From these halftone values, we would like to reconstruct the original image. Our eyes can do it, but it is not so easy to describe this ability in algorithmic terms.

The need for such a representation also comes from the need to manipulate the original image, e.g., rotate it or zoom on it. These operations are easy to perform on the original image, but it is not clear how to directly perform them on a halftone image. Instead, we reconstruct the original image, rotate it, and halftone the rotated image.

There exist inverse halftoning methods, but these methods are still far from optimal.

The problem is that halftoning loses information and is, therefore, a lossy compression. Hence, there may be several different images that lead to the same halftoned image. In the existing methods, we reverse the halftoning procedure by selecting one of such images. However, it may be beneficial to present not just a *single* possible original image, but the whole *range* of images that could lead to the given halftoned image. Thus, instead of the image in which the intensity $f(n_1, n_2)$ at every pixel has an exact value, we come up with an "interval-valued" image in which, at each pixel (n_1, n_2), we only know the interval $[\underline{f}(n_1, n_2), \overline{f}(n_1, n_2)]$ of possible values of intensity; this idea was explored in [8].

Acknowledgments

This work was supported in part by NSF grants EAR-0225670 and DMS-0532645 and by Texas Department of Transportation grant No. 0-5453.

The authors are thankful to the organizers of the International Conference on Fuzzy Mathematics and Its Applications, especially to Dr. S. B. Nimse, for their encouragements.

References

[1] M. G. Averill, G. Xiang, V. Kreinovich, G. R. Keller, S. A. Starks, P. S. Debroux, and J. Boehm, How to Reconstruct the Original Shape of a Radar Signal?, *Proceedings of the 24th International Conference of the North American Fuzzy Information Processing Society NAFIPS'2005*, Ann Arbor, Michigan, June 22-25, 2005, pp. 717–721.

[2] M. G. Averill, K. C. Miller, G. R. Keller, V. Kreinovich, R. Araiza, and S. A. Starks, Using Expert Knowledge in Solving the Seismic Inverse Problem, *International Journal of Approximate Reasoning* (to appear).

[3] G. Bardossy and J. Fodor, *Evaluation of Uncertainties and Risks in Geology*, Springer Verlag, Berlin, 2004.

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

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

[6] D. Berleant 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.

[7] D. Berleant, M.-P. Cheong, C. Chu, Y. Guan, A. Kamal, G. Sheblé, S. Ferson, and J. F. Peters, Dependable handling of uncertainty, *Reliable Computing* 9(6) (2003), pp. 407–418.

[8] S. D. Cabrera, K. Iyer, G. Xiang, and V. Kreinovich, On Inverse Halftoning: Computational Complexity and Interval Computations, *Proceedings of the 39th Conference on Information Sciences and Systems CISS'2005*, John Hopkins University, March 16-18, 2005, Paper 164.

[9] Th. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge, MA, 2001.

[10] E. Dantsin, V. Kreinovich, A. Wolpert, and G. Xiang, Population Variance Under Interval Uncertainty: A New Algorithm, *Reliable Computing*, 2006, Vol. 12, No. 4, pp. 273–280.

[11] E. Dantsin, A. Wolpert, M. Ceberio, G. Xiang, and V. Kreinovich, Detecting Outliers under Interval Uncertainty: A New Algorithm Based on Constraint Satisfaction, *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU'06*, Paris, France, July 2-7, 2006, pp. 802–809.

[12] P. Debroux, J. Boehm, F. Modave, V. Kreinovich, G. Xiang, J. Beck, K. Tupelly, R. Kandathi, L. Longpre, and K. Villaverde, Using 1-D Radar Observations to Detect a Space Explosion Core Among the Explosion Fragments: Sequential and Distributed Algorithms, *Proceedings of the 11th IEEE Digital Signal Processing Workshop*, Taos, New Mexico, August 1-4, 2004, pp. 273–277.

[13] R. Demicco and J. Klir, Eds., *Fuzzy Logic in Geology*, Academic Press, 2003.

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

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

[16] J. A. Hole, Nonlinear High-Resolution Three-Dimensional Seismic Travel Time Tomography. *J. Geophysical Research*, 1992, Vol. 97, pp. 6553–6562.

[17] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter, *Applied interval analysis: with examples in parameter and state estimation, robust control and robotics*, Springer-Verlag, London, 2001.

[18] G. Klir and B. Yuan, *Fuzzy sets and fuzzy logic: theory and applications.* Prentice Hall, Upper Saddle River, New Jersey, 1995.

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

[20] V. Kreinovich 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, *Proceedings of the Conference on Computability and Complexity in Analysis CCA'2003*, Cincinnati, Ohio, USA, August 28–30, 2003, pp. 19–54.

[21] V. Kreinovich, L. Longpré, S. A. Starks, G. Xiang, J. Beck, R. Kandathi, A. Nayak, S. Ferson, and J. Hajagos, "Interval Versions of Statistical Techniques, with Applications to Environmental Analysis, Bioinformatics, and Privacy in Statistical Databases", *Journal of Computational and Applied Mathematics*, 2007, Vol. 199, No. 2, pp. 418–423.

[22] V. Kreinovich, G. Xiang, S. A. Starks, L. Longpré, M. Ceberio, R. Araiza, J. Beck, R. Kandathi, A. Nayak, R. Torres, and J. Hajagos, "Towards combining probabilistic and interval uncertainty in engineering calculations: algorithms for computing statistics under interval uncertainty, and their computational complexity", *Reliable Computing*, 2006, Vol. 12, No. 6, pp. 471–501.

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

[24] W. A. Lodwick 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.

[25] M. Maceira, S. R. Taylor, C. J. Ammon, X. Yang, and A. A. Velasco, High-resolution Rayleigh wave slowness tomography of Central Asia. *Journal of Geophysical Research*, 2005, Vol. 110, paper B06304.

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

[27] H. T. Nguyen and E. A. Walker, *A first course in fuzzy logic*, CRC Press, Boca Raton, Florida, 2005.

[28] M. Nikravesh, "Soft computing-based computational intelligence for reservoir characterization", *Expert Syst. Appl.*, 2004, Vol. 26, No. 1, pp. 19–38.

[29] P. Nivlet, 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.

[30] P. Nivlet, 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.

[31] M. Orshansky, W. Wang, M. Ceberio, and G. Xiang, Interval-based Robust Statistical Techniques for Non-negative Convex Functions, with Application to Timing Analysis of Computer Chips, *Proceedings of the ACM Symposium on Applied Computing SAC'06*, Dijon, France, April 23-27, 2006, pp. 1629–1633.

[32] M. Orshansky, W. Wang, G. Xiang, and V. Kreinovich, Interval-based Robust Statistical Techniques for Non-negative Convex Functions with Application to Timing Analysis of Computer Chips, *Proceedings of the Second International Workshop on Reliable Engineering Computing*, Savannah, Georgia, February 22-24, 2006, pp. 197–212.

[33] R. Osegueda, 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

[34] S. Rabinovich, *Measurement Errors: Theory and Practice*, American Institute of Physics, New York, 2005.

[35] H. Regan, S. Ferson, and D. Berleant, Equivalence of five methods for bounding uncertainty, *Journal of Approximate Reasoning*, 2004, 36(1):1–30.

[36] N. C. Rowe, 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.

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

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

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

[40] G. Xiang, Fast algorithm for computing the upper endpoint of sample variance for interval data: case of sufficiently accurate measurements, *Reliable Computing*, 2006, Vol. 12, No. 1, pp. 59–64.

[41] G. Xiang, M. Ceberio, and V. Kreinovich, *Computing Population Variance and Entropy under Interval Uncertainty: Linear-Time Algorithms*, University of Texas at El Paso, Department of Computer Science, Technical Report UTEP-CS-06-28b, 2006, available as http://www.cs.utep.edu/vladik/2006/tr06-28b.pdf

[42] G. Xiang, 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.

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