10-2007

# Fast Algorithms for Computing Statistics under Interval Uncertainty: An Overview

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

Gang Xiang



Comments:
Technical Report: UTEP-CS-07-49a
In: Van-Nam Huynh, Yoshiteru Nakamori, Hiroakira Ono, Jonathan Lawry, Vladik Kreinovich, and Hung T. Nguyen (eds.), *Interval/Probabilistic Uncertainty and Non-Classical Logics*, Springer-Verlag, Berlin-Heidelberg-New York, 2008, pp. 19-31.

# Fast Algorithms for Computing Statistics under Interval Uncertainty: An Overview

Vladik Kreinovich[1] and Gang Xiang[1,2]

[1] University of Texas at El Paso, El Paso, TX 79968, USA vladik@utep.edu
[2] Philips Healthcare Informatics gxiang@acm.org

**Summary.** In many areas of science and engineering, it is desirable to estimate statistical characteristics (mean, variance, covariance, etc.) under interval uncertainty. For example, we may want to use the measured values $x(t)$ of a pollution level in a lake at different moments of time to estimate the average pollution level; however, we do not know the exact values $x(t)$ – e.g., if one of the measurement results is 0, this simply means that the actual (unknown) value of $x(t)$ can be anywhere between 0 and the detection limit DL. We must therefore modify the existing statistical algorithms to process such interval data.

Such a modification is also necessary to process data from statistical databases, where, in order to maintain privacy, we only keep interval ranges instead of the actual numeric data (e.g., a salary range instead of the actual salary).

Most resulting computational problems are NP-hard – which means, crudely speaking, that in general, no computationally efficient algorithm can solve all particular cases of the corresponding problem. In this paper, we overview practical situations in which computationally efficient algorithms exist: e.g., situations when measurements are very accurate, or when all the measurements are done with one (or few) instruments.

## 1 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 $V(x_1, \ldots, x_n) = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E)^2$.

*Comment.* The population variance is often computed by using an alternative formula $V = M - E^2$, where $M = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2$ is the population second moment.

*Comment.* In many practical situations, we are interested in an *unbiased* estimate of the population variance $V_u(x_1, \ldots, x_n) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - E)^2$. In this dissertation, we will describe how to estimate $V$ under interval uncertainty; since $V_u = \frac{n}{n-1} \cdot V$, we can easily transform estimates for $V$ into estimates for $V_u$.

## 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., [38]. In other words, often, we know the result $\widetilde{x}$ of measuring the desired characteristic $x$, and we know the upper bound $\Delta$ 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 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 [34, 35]. Detecting outliers is also important in bioinformatics [41].

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., [48].

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., [23].

## 3 Estimating Statistics Under Interval Uncertainty: A Problem

In all such cases, 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 solving in this dissertation.

**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 [30, 42]; for further developments, see, e.g., [4, 5, 6, 7, 16, 19, 32, 33, 39, 40, 43] and references therein.

## 4 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 $\overline{E} = \dfrac{1}{n} \cdot (\overline{x}_1 + \ldots + \overline{x}_n)$.

## 5 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 [17] which means, crudely speaking, that the worst-case computation time grows exponentially with $n$. Specifically, computing the upper endpoint $\overline{V}$ of the range $[\underline{V}, \overline{V}]$ is NP-hard. Moreover, if we want to compute the variance range or $\overline{V}$ with a given accuracy $\varepsilon$, the problem is still NP-hard. (For a more detailed description of NP-hardness in relation to interval uncertainty, see, e.g., [22].)

## 6 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 $\Delta x_i$, and these errors are small, we can ignore terms that are quadratic (and of higher order) in $\Delta x_i$ 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 $\Delta x_i$ can take all possible values from $-\Delta_i$ to $\Delta_i$, 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 $\Delta \stackrel{\text{def}}{=} \sum\limits_{i=1}^{n} |C_i| \cdot \Delta_i$.

In particular, if we take, as the statistic, the population variance $C = V$, then $C_i = \dfrac{\partial V}{\partial x_i} = \dfrac{2}{n} \cdot (\widetilde{x}_i - \widetilde{E})$, where $\widetilde{E}$ is the average of the midpoints $\widetilde{x}_i$,

and $C_0 = \frac{1}{n} \cdot \sum_{i=1}^{n} (\widetilde{x}_i - \widetilde{E})^2$ is the variance of the midpoint values $\widetilde{x}_1, \ldots, \widetilde{x}_n$. So, for the variance, $\Delta = \frac{2}{n} \cdot \sum_{i=1}^{n} |\widetilde{x}_i - \widetilde{E}| \cdot \Delta_i$.

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

## 7 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.

## 8 First Class: Narrow Intervals

The main idea behind linearization is that if the measurement errors $\Delta x_i$ are small, we can safely ignore quadratic and higher order terms in $\Delta x_i$ and replace the original difficult-to-analyze expression by its easier-to-analyze linear approximation. The accuracy of this techniques is determined by the size of the first term that we ignore, i.e., is of size $O(\Delta x_i^2)$. Thus, the narrower the intervals (i.e., the smaller the values $\Delta x_i$), the more accurate is the result of this linearization.

In real life, we want to compute the range with a certain accuracy. So, when the intervals are sufficiently accurate, the results of linearization estimation provide the desired accuracy and thus, we have a feasible algorithm for solving our problem. When the intervals become wider, we can no longer ignore the quadratic terms and thus, the problem becomes more computationally complex. In other words, when intervals are narrower, the problem of computing statistics under interval uncertainty becomes easier. It is therefore

reasonable to consider the case of narrow intervals as the first case in which we can expect feasible algorithms for computing statistics of interval data.

How can we describe "narrowness" formally? The very fact that we are performing the statistical analysis means that we assume that the actual values $x_1, \ldots, x_n$ come from a probability distribution, and we want to find the statistical characteristics of this probability distribution. Usually, this distribution is continuous: normal, uniform, etc. Formally, a continuous distribution is a one for which a finite probability density $\rho(x)$ exists for every $x$. In this case, for every the real number $a$, the probability $p = \int_{a-\delta}^{a+\delta} \rho(x)\,dx$ to have a random value within an interval $[a-\delta, a+\delta]$ is approximately equal to $\rho(a) \cdot 2\delta$ and thus, tends to 0 as $\delta \to 0$. This means that for every value $a$, the probability to have a random value exactly equal to $a$ is 0. In particular, this means that with probability 1, all the values $x_1, \ldots, x_n$ randomly selected from the original distribution are different.

The data intervals $\mathbf{x}_1, \ldots, \mathbf{x}_n$ contain these different values $x_1, \ldots, x_n$. 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.

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

## 9 Second Class: Slightly Wider Intervals

Narrow intervals can be described as intervals which do not intersect at all. Namely, we have a set of (unknown) actual values $x_1 < x_2 < \ldots < x_n$, and we have intervals around each value which are so narrow that the neighboring intervals $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$ do not intersect.

As the widths of the intervals increase, they start intersecting. At first, only the neighboring intervals $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$ intersect, but intervals $\mathbf{x}_i$ and $\mathbf{x}_{i+2}$ still do not intersect. As the widths increase further, intervals $\mathbf{x}_i$ and $\mathbf{x}_{i+2}$ start intersecting, etc. When the intervals become very wide, all $n$ intervals intersect.

We can therefore gauge the degree of narrowness by the number of intervals which have a common point.

Specifically, we define the case of slightly wider intervals as the situation when for some integer $K$, no set of $K$ intervals has a common intersection. The case of narrow intervals correspond to $K = 2$, the next case is $K = 3$, etc. – all the way to the general case $K = n$.

As we have mentioned, the narrower the intervals, the easier the corresponding computational problem. Since the parameter $K$ is a measure of this narrowness, it is therefore reasonable to expect that feasible algorithms exist in this case – at least for values of $K$ which are not too large.

## 10 Third Class: Single Measuring Instrument

We have already mentioned that one of the most widely used engineering techniques for dealing with measurement uncertainty is linearlization. To be able to easily compute the range $\mathbf{C}$ of a statistic $C$ by using linearization, we must make sure not only that intervals are relatively narrow, but also that they are approximately of the same size: otherwise, if, say, $\Delta x_i^2$ is of the same order as $\Delta x_j$, we cannot meaningfully ignore $\Delta x_i^2$ and retain $\Delta x_j$. 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).

The traditional linearization techniques only provide us with an approximate range. However, as we will show, for some classes of problems, these approximate estimates can be refined into an efficient computation of the exact range. Because of this possibility, let us formulate, in precise terms, the class of problems for which linearization is possible, i.e., the class of problem for which all the measurements have been performed by a single measuring instrument.

How can we describe this class 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: $[\underline{x}_i, \overline{x}_i] \subseteq (\underline{x}_j, \overline{x}_j)$.

This restriction only refers to not absolutely measurement results, i.e., to non-degenerate intervals. In addition to such interval values, we may also have machine-represented floating point values produced by very accurate measurements, so accurate that we can, for all practical purposes, consider these values exactly known. 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.

As we will see, the absence of such pairs is a useful property that enables us to compute interval statistics faster. We will also see that this absence happens not only for measurements made by a single MI, but also in several other useful practical cases. Since this property is useful, we will give it a name.

We 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.

## 11 Fourth Class: Several MI

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

We have already mentioned that the case of a single MI is the easiest; the more MI we involve, the more complex the resulting problem – all the way to

the general case $m = n$, when each measurement is performed by a different MI.

Since the parameter $m$ is a measure of complexity, it is therefore reasonable to expect that feasible algorithms exist for the case of a fixed number $m$ – at least for the values of $m$ which are not too large.

## 12 Fifth Class: Privacy Case

In the previous text, we mainly emphasized that measurement uncertainty naturally leads to intervals. It is worth mentioning, however, that interval uncertainty may also come from other sources: e.g., from the desire to protect privacy in statistical databases. Indeed, often, we collect large amounts of data about persons – e.g., during census, or during medical experiments. Statistical analysis of this data enables us to find useful correlations between, e.g., age and effects of a certain drug, or between a geographic location and income level. Because of this usefulness, it is desirable to give researchers an ability to perform a statistical analysis of this data. However, if we simply researchers to receive answers to arbitrary queries and publish the results of their analysis, then these results may reveal the information from the databases which is private and not supposed to be disclosed.

One way to protect privacy is not to keep the exact actual values of the privacy-related quantities such as salary or age in the database. Instead, we fix a finite number of thresholds, e.g., 0, 10, 20, 30 years, and for each person, we only record the corresponding age range: from 0 to 10, or from 10 to 20, or from 20 to 30, etc. Since the actual values are not stored in the database anymore, no queries can disclose these values.

So, this idea solves the privacy problem, but it opens up another problem: how can perform statistical processing on this privacy-related interval data? Suppose that we are interested in the values of a statistical characteristic $C(x_1, \ldots, x_n)$. If we knew the actual values $x_1, \ldots, x_n$, then we could easily compute the value of this characteristic. However, in case of privacy-related interval uncertainty, all we know is intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ of possible values of $x_i$. Different values $x_i \in \mathbf{x}_i$ lead, in general, to different values of $C(x_1, \ldots, x_n)$. So, a reasonable idea is to return the range of possible values of the characteristic $C(x_1, \ldots, x_n)$ when $x_i \in \mathbf{x}_i$.

From the algorithmic viewpoint, we get the same problem as with measurement-related interval uncertainty: find the range of the given characteristic $C(x_1, \ldots, x_n)$ on given intervals $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The *difference* between this case and the two previous cases is that, in the first two cases, we *do not know the exact values*, while in this case, in principle, it is *possible to get the exact value*, but we do not use the exact values, because we want to protect privacy.

From the mathematical viewpoint, privacy-related intervals have the following property: they either coincide (if the value corresponding to the two

patients belongs to the same range) or are different, in which case they intersect in at most 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 in at most one point – *privacy case.*

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

*Comment.* Sometimes, in the privacy-motivated situation, we must process interval data in which intervals come from several different "granulation" schemes. For example, to find the average salary in North America, we may combine US interval records in which the salary is from 0 to 10,000 US dollars, from 10,000 to 20,000, etc., with the Canadian interval records in which the ranges are between 0 to 10,000 Canadian dollars, 10,000 to 20,000 Canadian dollars, etc. When we transform these records to a single unit, we get two different families of intervals, each of which satisfies the subset property. Thus, to handle such situations, we can use algorithms developed for the several MI case.

## 13 Sixth Class: Non-Detects

An important practical case is the case of non-detects. Namely, many sensors are reasonably accurate, but they have a detection limit $DL$ – so they cannot detect any value below $DL$ but they detect values of $DL$ and higher with a very good accuracy.

In this case, if a sensor returns a value $\widetilde{x} \geq DL$, then this value is reasonably accurate, so we can consider it exact (i.e., a degenerate interval $[\widetilde{x}, \widetilde{x}]$). However, if the sensor does not return any signal at all, i.e., the measurement result $\widetilde{x} = 0$, then the only thing we can conclude about the actual value of the quantity is that this value is below the detection limit, i.e., that it lies in the interval $[0, DL]$.

In this case, every interval 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). Under this assumption, the resulting non-degenerate intervals also satisfy the subset property. Thus, algorithms that work for the subset property 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 $DL$.

## 14 Results

The main results are summarized in the following table:

| Case | $E$ | $V$ | $L, U$ | $S$ |
|---|---|---|---|---|
| Narrow intervals | $O(n)$ | $O(n)$ | $O(n \cdot \log(n))$ | $O(n^2)$ |
| Slightly wider narrow intervals | $O(n)$ | $O(n \cdot \log(n))$ | $O(n \cdot \log(n))$ | ? |
| Single MI | $O(n)$ | $O(n)$ | $O(n \cdot \log(n))$ | $O(n^2)$ |
| Several ($m$) MI | $O(n)$ | $O(n^m)$ | $O(n^m)$ | $O(n^{2m})$ |
| New case | $O(n)$ | $O(n^m)$ | ? | ? |
| Privacy case | $O(n)$ | $O(n)$ | $O(n \cdot \log(n))$ | $O(n^2)$ |
| Non-detects | $O(n)$ | $O(n)$ | $O(n \cdot \log(n))$ | $O(n^2)$ |
| General | $O(n)$ | NP-hard | NP-hard | ? |

**Table 1.** Computational complexity of statistical analysis under interval uncertainty: an overview

Here, $E$ is a population mean, $V$ is a population variance, $S \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E)^3$ is the population skewness, and $L \stackrel{\text{def}}{=} E - k_0 \cdot \sigma$ and $U \stackrel{\text{def}}{=} E + k_0 \cdot \sigma$ are endpoints of the confidence interval, where a parameter $k_0$ is usually taken as $k_0 = 2$, $k_0 = 3$, or $k_0 = 6$.

*Comment.* For descriptions of the algorithms, and for proofs of the algorithm correctness, see [18, 46] and references therein; see also [1, 3, 12, 13, 14, 20, 21, 23, 24, 25, 26, 27, 28, 29, 31, 44, 45, 47].

**Applications.** There are several application areas in which it is possible to take into account interval uncertainty in statistical data processing:

- the seismic inverse problem in geophysics [2],
- the problem of estimating and decreasing the clock cycle in computer chips [36, 37],
- the problem of separating the core from the fragments in radar data processing [15], and
- the problem of inverse half-toning in image processing [11].

## 15 Conclusion

In many areas of science and engineering, it is desirable to estimate statistical characteristics (mean, variance, covariance, etc.) under interval uncertainty. Such a modification is necessary, e.g., to process data from statistical databases, where, in order to maintain privacy, we only keep interval ranges instead of the actual numeric data (e.g., a salary range instead of the actual salary).

Most resulting computational problems are NP-hard – which means, crudely speaking, that in general, no computationally efficient algorithm can solve all particular cases of the corresponding problem.

In this paper, we overview practical situations in which computationally efficient algorithms exist: e.g., situations when measurements are very accurate, or when all the measurements are done with one (or few) instruments.

## Acknowledgments

This work was supported in part by NSF grants HRD-0734825, EAR-0225670, and EIA-0080940, by Texas Department of Transportation grant No. 0-5453, by the Max Planck Institut für Mathematik, and by the Japan Advanced Institute of Science and Technology (JAIST) International Joint Research Grant 2006-08.

The authors are thankful to Dr. Van Nam Huynh for valuable suggestions.

## References

1. R. Alo, M. Beheshti, and G. Xiang, Computing variance under interval uncertainty: a new algorithm and its potential application to privacy in statistical databases, *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU'06*, Paris, France, Jul. 2–7, 2006, pp. 810–816.
2. M. G. Averill, K. C. Miller, G. R. Keller, V. Kreinovich, R. Araiza, S. A. Starks, Using expert knowledge in solving the seismic inverse problem, *International Journal of Approximate Reasoning*, 2007, Vol. 45, No. 3, pp. 564–587.
3. J. Beck, 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 (eds.), *Soft Methodology and Random Information Systems*, Springer-Verlag, Berlin-Heidelberg, 2004, pp. 85–92.
4. D. Berleant, Automatically verified arithmetic with both intervals and probability density functions, *Interval Computations*, 1993, No. 2, pp. 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, Vol. 4, No. 2, pp. 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*, 2003, Vol. 9, No. 6, pp. 407–418.
8. D. Berleant, 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, Vol. 9, No. 2, pp. 91–108.
9. D. Berleant and J. Zhang, Using Pearson correlation to improve envelopes around the distributions of functions, *Reliable Computing*, 2004, Vol. 10, No. 2, pp. 139–161.

10. D. Berleant and J. Zhang, Representation and problem solving with the distribution envelope determination (DEnv) method, *Reliability Engineering and System Safety*, Jul.–Sep. 2004, Vol. 85, No.1–3.
11. 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, Mar. 16–18, 2005, Paper 164.
12. M. Ceberio, S. Ferson, V. Kreinovich, S. Chopra, G. Xiang, A. Murguia, and J. Santillan, How to take into account dependence between the inputs: from interval computations to constraint-related set computations, with potential applications to nuclear safety, bio- and geosciences, *Proceedings of the Second International Workshop on Reliable Engineering Computing*, Savannah, Georgia, Feb. 22–24, 2006, pp. 127–154.
13. 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.
14. 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, Jul. 2-7, 2006, pp. 802–809.
15. P. Debroux, J. Boehm, F. Modave, V. Kreinovich, G. Xiang, J. Beck, K. Tupelly, R. Kandathi, L. Longpré, 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, Aug. 1-4, 2004, pp. 273–277.
16. S. Ferson, *RAMAS Risk Calc 4.0: Risk Assessment with Uncertain Numbers*, CRC Press, Boca Raton, Florida, 2002.
17. S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, Exact bounds on finite populations of interval data, *Reliable Computing*, 2005, Vol. 11, No. 3, pp. 207–233.
18. S. Ferson, V. Kreinovich, J. Hajagos, W. Oberkampf, and L. Ginzburg, *Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty*, Sandia National Laboratories, Report SAND2007-0939, May 2007.
19. S. Ferson, D. Myers, and D. Berleant, *Distribution-Free Risk Analysis: I. Range, Mean, and Variance*, Applied Biomathematics, Technical Report, 2001.
20. L. Granvilliers, V. Kreinovich, and N. Müller, Novel approaches to numerical software with result verification, In: R. Alt, A. Frommer, R. B. Kearfott, and W. Luther (eds.), *Numerical Software with Result Verification*, (International Dagstuhl Seminar, Dagstuhl Castle, Germany, Jan. 19–24, 2003), Springer Lectures Notes in Computer Science, 2004, Vol. 2991, pp. 274–305.
21. V. Kreinovich, Probabilities, intervals, what next? optimization problems related to extension of interval computations to situations with partial information about probabilities, *Journal of Global Optimization*, 2004, Vol. 29, No. 3, pp. 265–280.
22. V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer, Dordrecht, 1997.
23. 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, editors, *Proc. Conf. on Computability and Complexity in Analysis CCA'2003*, Cincinnati, Ohio, USA, Aug. 28–30, 2003, pp. 19–54.
24. V. Kreinovich, L. Longpré, P. Patangay, S. Ferson, and L. Ginzburg, Outlier detection under interval uncertainty: algorithmic solvability and computational complexity, *Reliable Computing*, 2005, Vol. 11, No. 1, pp. 59–76.
25. 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.
26. V. Kreinovich, 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*, 2007, Vol. 177, No. 16, pp. 3228–3238.
27. V. Kreinovich, G. N. Solopchenko, S. Ferson, L. Ginzburg, and R. Aló, Probabilities, intervals, what next? extension of interval computations to situations with partial information about probabilities, *Proceedings of the 10th IMEKO TC7 International Symposium on Advances of Measurement Science*, St. Petersburg, Russia, Jun. 30–Jul. 2, 2004, Vol. 1, pp. 137–142.
28. V. Kreinovich, G. Xiang, and S. Ferson, Efficient algorithms for computing mean and variance under dempster-shafer uncertainty, *International Journal of Approximate Reasoning*, 2006, Vol. 42, pp. 212–227.
29. 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.
30. V. P. Kuznetsov, *Interval Statistical Models*, Radio i Svyaz, Moscow, 1991 (in Russian).
31. A. T. Langewisch and F. F. Choobineh, Mean and variance bounds and propagation for ill-specified random variables, *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 2004, Vol. 34, No. 4, pp. 494–506.
32. 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, Vol. 9, No. 2, pp. 127–141.
33. R. E. Moore and W. A. Lodwick, Interval analysis and fuzzy set theory, *Fuzzy Sets and Systems*, 2003, Vol. 135, No. 1, pp. 5–9.
34. 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, Sep. 9–14, 2001, pp. 1644–1647.
35. 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, Sep. 30–Oct. 3, 2001, paper SPE-71327.
36. 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, Apr. 23-27, 2006, pp. 1629–1633.

37. 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, Feb. 22–24, 2006, pp. 197–212.
38. S. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice*, Springer-Verlag, New York, 2005.
39. H. Regan, S. Ferson, and D. Berleant, Equivalence of five methods for bounding uncertainty, *Journal of Approximate Reasoning*, 2004, Vol. 36, No. 1, pp. 1–30.
40. 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, Vol. 9, pp. 1098–1113.
41. I. Shmulevich and W. Zhang, Binary analysis and optimization-based normalization of gene expression data, *Bioinformatics*, 2002, Vol. 18, No. 4, pp. 555–565.
42. P. Walley, *Statistical Reasoning with Imprecise Probabilities*, Chapman & Hall, N.Y., 1991.
43. R. Williamson and T. Downs, Probabilistic arithmetic I: numerical methods for calculating convolutions and dependency bounds, *International Journal of Approximate Reasoning*, 1990, Vol. 4, pp. 89–158.
44. G. Xiang, S. A. Starks, V. Kreinovich, and L. Longpré, New Algorithms for statistical analysis of interval data, In: J. Dongarra, K. Madsen, and J. Wasniewski (eds.), *PARA'04 Workshop on State-of-the-Art in Scientific Computing*, Springer Lecture Notes in Computer Science, 2005, Vol. 3732, pp. 189–196.
45. 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.
46. G. Xiang, *Fast Algorithms for Computing Statistics under Interval Uncertainty, with Applications to Computer Science and to Electrical and Computer Engineering*, PhD Dissertation, Computer Science Dept., University of Texas at El Paso, July 2007.
47. G. Xiang, O. Kosheleva, and G. J. Klir, Estimating information amount under interval uncertainty: algorithmic solvability and computational complexity, *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU'06*, Paris, France, Jul. 2–7, 2006, pp. 840–847.
48. W. Zhang, I. Shmulevich, and J. Astola, *Microarray Quality Control*, Wiley, Hoboken, New Jersey, 2004.