Efficient Algorithms for Computing Mean and Variance Under Dempster-Shafer Uncertainty

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

[1]Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA, vladik@utep.edu
[2]Applied Biomathematics, 100 North Country Road
Setauket, NY 11733, USA, scott@ramas.com

Abstract

In many real-life situations, we only have partial information about the actual probability distribution. For example, under Dempster-Shafer uncertainty, we only know the masses $m_1, \ldots, m_n$ assigned to different sets $S_1, \ldots, S_n$, but we do not know the distribution within each set S_i. Because of this uncertainty, there are many possible probability distributions consistent with our knowledge; different distributions have, in general, different values of standard statistical characteristics such as mean and variance. It is therefore desirable, given a Dempster-Shafer knowledge base, to compute the *ranges* of possible values of mean and of variance.

The existing algorithms for computing the range for the variance require $\approx 2^n$ computational steps, and therefore, cannot be used for large n. In this paper, we propose new efficient algorithms that work for large n as well.

1 Formulation of the Problem

In many real-life situations, we only have partial information about the actual probability distribution. In many practical situations, this uncertainty is naturally described by a Dempster-Shafer (DS) approach (see, e.g., [14]), in which the knowledge consists of a finite collection of sets $S_1, \ldots, S_n$ and non-negative "masses" (probabilities) $m_1, \ldots, m_n$ assigned to these sets in such a way that $m_1 + \ldots + m_n = 1$,

In particular, in the 1-D case, instead of the exact probability distribution, we have a finite collection of intervals $\mathbf{x}_1 = [\underline{x}_1, \overline{x}_1]$, $\ldots$, $\mathbf{x}_n = [\underline{x}_n, \overline{x}_n]$, and we have non-negative "masses" (probabilities) $m_1, \ldots, m_n$ assigned to these intervals in such a way that $m_1 + \ldots + m_n = 1$.

Definition 1. *By a (1-D)* Dempster-Shafer knowledge base*, we mean a pair*

$$K = \langle (\mathbf{x}_1, \ldots, \mathbf{x}_n), (m_1, \ldots, m_n) \rangle, \tag{1}$$

where $\mathbf{x}_i$ *are intervals and* m_i *are positive numbers for which* $\sum m_i = 1$.

Let us recall how the corresponding knowledge base is interpreted in probabilistic terms. In the simplest case when the Dempster-Shafer knowledge base consists of a single interval $\mathbf{x}_1$ with the mass $m_1 = 1$, this means that we are sure that the actual probability distribution with the probability density $\rho_1(x)$ is located on this interval, i.e., $\rho_1(x) = 0$ for $x \notin \mathbf{x}_1$, but we do not know what exactly distribution we have (i.e., we do not know the exact probability density $\rho_1(x)$).

If we have several intervals $\mathbf{x}_i$, this means that:

- with probability m_1, we select the interval $\mathbf{x}_1$,
- with probability m_2, we select the interval $\mathbf{x}_2$,
- ...
- with probability m_n, we select the interval $\mathbf{x}_n$.

Then, within the selected interval $\mathbf{x}_i$, we select a value x according to some probability distribution $\rho_i(x)$ located on this interval. As a result, the overall probability distribution takes the form

$$\rho(x) = m_1 \cdot \rho_1(x) + \ldots + m_n \cdot \rho_n(x). \tag{2}$$

So, the original Dempster-Shafer knowledge base means that the actual (unknown) probability distribution is of the above type, with $\rho_i(x)$ located on the interval $\mathbf{x}_i$.

Definition 2. *Let* K *be a Dempster-Shafer knowledge base described by the formula (1). We say that a probability distribution* $\rho(x)$ *is* consistent *with* K *if it has the form (2) for some probability distributions* $\rho_i(x)$ *located on the intervals* $\mathbf{x}_i$.

Comment. For some probability distributions, there is no probability density function: e.g., for a probability distribution that is located at a point x_0 with probability 1. In this case, instead of continuous functions $\rho(x)$, we also allow "functions" $\rho(x)$ defined as limits of continuous functions – with appropriately defined limits as integrals. Such limits are called *generalized functions*, or *distributions*; see, e.g., [2]. For example, the above degenerate probability distribution can be described by a *delta-function* density $\rho(x) = \delta(x - x_0)$.

There exist infinitely many probability distributions of the above type (2). For each of these distributions ρ, we can find the value of the corresponding statistical characteristic $C(\rho)$ – e.g., the mean $E = \int x \cdot \rho(x)\,dx$ or the variance $V = \int (x - E)^2 \cdot \rho(x)\,dx$. As a natural extension of the original probability distribution to the Dempster-Shafer knowledge base, we can thus take the range of values of the characteristic $C(\rho)$ among all probability distributions $\rho(x)$ which are consistent with this knowledge base, i.e.,

$$C(K) \stackrel{\text{def}}{=}$$

$$\left\{ C(\rho) : \rho(x) = \sum_{i=1}^{n} m_i \cdot \rho_i(x) \text{ for some distributions } \rho_i(x) \text{ located on } \mathbf{x}_i \right\}.$$

Definition 3. *By a* statistical characteristic*, we mean a mapping which assigns, to every probability distribution ρ, a real number $C(\rho)$.*

Definition 4. *Let $C(\rho)$ be a statistical characteristic, and let K be a Dempster-Shafer knowledge base. By a* range *$C(K)$ of the characteristic C on the knowledge base K, we mean the set of all the values $C(\rho)$ when ρ is consistent with K.*

A natural question is: how can we compute this range for such natural statistical characteristics as mean and variance?

Algorithms for computing these ranges are known; see, e.g., [11]. However, the number of computational steps which are needed for some of these algorithms grows exponentially (as $\sim 2^n$) with the size n of the knowledge base. As a result, while it is quite possible to compute the exact range when n is small (e.g., $n \approx 10$), for larger n (e.g., for $n \approx 100$), these algorithms are no longer feasible.

In this paper, we produce a new algorithm that computes these ranges in *feasible* time – i.e., in time that grows polynomially with the size of the problem.

2 A Similar (But Different) Problem: Computing Mean and Variance Under Interval Uncertainty

Formulation of the similar problem. A similar problem arises in the related situation of *interval* uncertainty. This similar problem is related to the following natural question: if, instead of the actual distribution, we only have a sample $x_1, \ldots, x_n$ from this distribution, how can we estimate the mean and variance of the distribution? In practice, the most widely used estimated are the population mean $E = \dfrac{x_1 + \ldots + x_n}{n}$ and the population average $V = \dfrac{(x_1 - E)^2 + \ldots + (x_n - E)^2}{n}$; see, e.g., [13].

The values x_i usually come from measurement, and measurement results are never absolutely accurate; the measured values $\widetilde{x}_i$ usually slightly differ from the true (unknown) values x_i of the measured quantities. Often, the only information that we have about the measurement error $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$ is the upper bound Δ_i provided by the manufacturer of the corresponding measuring instrument; see, e.g., [12]. In this situation, the only information that we have about the true (unknown) value x_i is that x_i belongs to the interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$, where $\underline{x}_i = \widetilde{x}_i - \Delta_i$ and $\overline{x}_i = \widetilde{x}_i + \Delta_i$.

For different values $x_i \in \mathbf{x}_i$, we get, in general, different values of mean and variance. It is therefore desirable, given n intervals $\mathbf{x}_1, \ldots, \mathbf{x}_n$, to compute the range

$$\mathbf{E} \stackrel{\text{def}}{=} \{E(x_1, \ldots, x_n) : x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}$$

of possible values of the population mean E and the range

$$\mathbf{V} \stackrel{\text{def}}{=} \{V(x_1, \ldots, x_n) : x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}$$

of possible values of the population variance V.

Why the interval problem is similar to our DS problem. In the interval problem, we have n intervals, and we want to find the ranges for the mean and for the variance. In the particular case of the DS problem when all the masses are equal (i.e., $m_1 = \ldots = m_n = 1/n$), we also have n intervals $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and we are also interested in finding the ranges for the mean and for the variance.

Because of this similarity, it is reasonable to use the experience of solving the interval problem in solving our DS problem.

What is known about the interval problem. Since the population mean is a monotonic function of n variables x_i, its smallest possible value $\underline{E}$ is attained when all the values x_i are the smallest possible (i.e., when $x_i = \underline{x}_i$ for all i), and, correspondingly, its largest possible value $\overline{E}$ is attained when all the values x_i are the largest possible (i.e., when $x_i = \overline{x}_i$ for all i). Thus, the range $\mathbf{E} = [\underline{E}, \overline{E}]$ of the population average can be computed as follows:

$$\underline{E} = \frac{\underline{x}_1 + \ldots + \underline{x}_n}{n}, \quad \overline{E} = \frac{\overline{x}_1 + \ldots + \overline{x}_n}{n}.$$

For the variance $\mathbf{V} = [\underline{V}, \overline{V}]$, there exist efficient algorithms for computing $\underline{V}$ and – under some reasonable condition – of computing $\overline{V}$, but in general, the problem of computing $\overline{V}$ has been proven to be NP-hard; see, e.g., [3, 4, 6]. (Crudely speaking, NP-hard means that in general, we cannot compute the exact range $\mathbf{V}$ faster than in exponential time $\approx 2^n$.)

This NP-hardness result may sound somewhat discomforting. However, as we will show in this paper, the DS problem is different from the similar interval problem, and, because of this difference, we can modify the interval-related algorithms into efficient DS algorithms.

Why the DS problem is different from the interval problem. We will show that for variance, the interval range is, in general, different from the DS range corresponding to the case $m_1 = \ldots = m_n = 1/n$ – even for the case when we have a single interval $[\underline{x}_1, \overline{x}_1]$.

Indeed, if we have a single interval, then in the Dempster-Shafer case this means that we can have an arbitrary distribution located on this interval. One can check that in this case:

- the smallest possible of the variance is 0 – when this distribution is located on a single value $x_1 \in [\underline{x}_1, \overline{x}_1]$ with probability 1, and
- the largest possible value of the variance is equal to $(\overline{x}_1 - \underline{x}_1)^2/4$ – when the random variable is located at each of the endpoints with the probability 1/2.

Thus, in this case, $C(K) = [0, (\overline{x}_1 - \underline{x}_1)^2/4]$.

On the other hand, for a single value $x_1 \in [\underline{x}_1, \overline{x}_1]$, the population variance $\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E)^2$ is equal to 0 no matter what is the actual value $x_1 \in [\underline{x}_1, \overline{x}_1]$. Hence, the corresponding interval is equal to $[0, 0]$ – i.e., for the case when $\underline{x}_1 < \overline{x}_1$, the interval corresponding to the Dempster-Shafer case is different from the interval corresponding to the population statistics.

One might think that this difference is caused by the fact that we have a *single* interval. However, it is easy to come up with similar examples when we have *several* intervals with $m_1 = \ldots = m_n = 1/n$. For example, when $n = 3$ and $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}_3 = [0, 1]$, in the DS approach, it is possible that on each of these intervals, we have a distribution that is located on each endpoint with probability 1/2. In this case, we attain the variance $V = 1/4$ – the largest possible variance that we can attain for any probability distribution located on the interval $[0, 1]$.

If on each interval, we pick the same value 1/2 with probability 1, then the variance is 0. Since the variance is always non-negative, we conclude that, in the DS approach, $V(K) = [0, 1/4]$.

Let us now estimate the corresponding interval range. Since the population variance is a non-negative quadratic function, its maximum is attained when each of the variables takes one of the extreme values $\underline{x}_i = 0$ or $\overline{x}_i = 1$. Out of possible combinations, the population variance attains its largest value when the values of x_i are different, i.e., when two values coincide with 0 or 1, and the third value is equal to, correspondingly, 1 or 0. In this case, the largest possible value of population variance is

$$V = \frac{1}{3} \cdot \left(\left(\frac{2}{3} \right)^2 + 2 \cdot \left(\frac{1}{3} \right)^2 \right) = \frac{1}{3} \cdot \frac{6}{9} = \frac{2}{9},$$

which is smaller than 1/4.

3 First (Simple) Result: Computing Mean (and Other Monotonic Statistical Characteristics) Under Dempster-Shafer Uncertainty

For the mean $C = E$, the algorithm $\mathcal{E}$ for computing the DS range $E(K)$ is as follows: the range is $[\underline{E}, \overline{E}]$, where

$$\underline{E} = \sum_{i=1}^{n} m_i \cdot \underline{x}_i, \quad \overline{E} = \sum_{i=1}^{n} m_i \cdot \overline{x}_i.$$

Proposition 1. *The algorithm $\mathcal{E}$ always computes $E(K)$ in time $O(n)$.*

Proof. The mean E of an arbitrary distribution of the type (2) can be described as $\sum_{i=1}^{n} \int x \cdot \rho_i(x)\, dx$, where each i-th integral $\int x \cdot \rho_i(x)\, dx$ is over the i-th interval $[\underline{x}_i, \overline{x}_i]$. For each i, the corresponding integral is the smallest if x_i is the smallest, i.e., if $x_i = \underline{x}_i$ with probability 1. Similarly, for each i, the corresponding integral is the largest if x_i is the largest, i.e., if $x_i = \overline{x}_i$ with probability 1. Thus, the above formulas indeed describe the desired range for E.

Each of these two formulas requires n additions and one division, hence overall, we need $O(n)$ computational steps. The proposition is proven.

Comment. For the case when $m_1 = \ldots = m_n = 1/n$, these DS-related formulas coincide with the above formulas for the range of E under interval uncertainty.

It turns out that the same is true for all statistical characteristics which are monotonic (in some reasonable sense). To describe this definition formally, let us recall the notion of *stochastic dominance* – a natural generalization of standard order to probability distributions. Namely, if we know the exact values x and y of two variables, then we can say that y *dominates* x if $y \geq x$. If x and y are random variables, then it is natural to say that y *dominates* x if for every real number t, the probability that y exceeds t is larger than (or equal to) the probability that x exceeds t.

The probability $\mathrm{Prob}(x > t)$ that $x > t$ can be described as $1 - F_x(t)$, where $F_x(t) \stackrel{\mathrm{def}}{=} \mathrm{Prob}(x \leq t)$ is the corresponding value of the cumulative distribution function (cdf). Thus, the condition that $1 - F_y(t) \geq 1 - F_x(t)$ can be reformulated as $F_y(t) \leq F_x(t)$. So, we arrive at the following definition:

Definition 5. *We say that a probability distribution with a cumulative distribution function $F_y(t)$* dominates *a probability distribution with a cumulative distribution function $F_x(t)$ if $F_y(t) \leq F_x(t)$ for every real number t.*

Definition 6. *We say that a statistical characteristic is* monotonic *if $C(\rho) \geq C(\rho')$ whenever the distribution described by the density ρ dominates the distribution described by the density ρ'.*

Comment. Mean is a monotonic characteristic; another monotonic characteristic is the *median*, i.e., the value m for which $F(m) = 1/2$.

Proposition 2. *For every monotonic statistical characteristic C, the range $C(K)$ is equal to $[C(\underline{x}), C(\overline{x})]$, where $\underline{x}$ is a probability distribution in which we have $\underline{x}_i$ with probability m_i, and $\overline{x}$ is a probability distribution in which we have $\overline{x}_i$ with probability m_i.*

Comment. If, for several intervals $\mathbf{x}_i$, their lower endpoints $\underline{x}_i$ coincide, then, of course, we have to add the corresponding probabilities m_i to describe the probability of the corresponding lower endpoint; same for upper endpoints.

Variance is not monotonic: e.g., the degenerate distribution in which $x = 1$ with probability 1 dominates the uniform distribution on the interval $[0, 1]$, but the variance of the degenerate distribution is equal to 0 and is, hence, smaller than the variance of uniform distribution. Thus, for the variance V, we have to come up with new algorithms for computing the corresponding range $V(K) = [\underline{V}, \overline{V}]$.

4 Main Result: Computing Variance Under Dempster-Shafer Uncertainty

The algorithm $\underline{\mathcal{V}}$ for computing $\underline{V}$ is as follows:

- First, we sort all $2n$ values $\underline{x}_i$, $\overline{x}_i$ into a sequence $x_{(1)} < x_{(2)} < \ldots < x_{(q)}$ for some $q \le 2n$. We will take $x_{(q+1)} \stackrel{\text{def}}{=} +\infty$.
- Second, we use bisection to find the value k ($1 \le k \le q$) for which the following two inequalities hold:

$$\sum_{j:\overline{x}_j \le x_{(k)}} m_j \cdot (x_{(k)} - \overline{x}_j) \le \sum_{i:\underline{x}_i \ge x_{(k+1)}} m_i \cdot (\underline{x}_i - x_{(k)}); \tag{3}$$

$$\sum_{j:\overline{x}_j \le x_{(k)}} m_j \cdot (x_{(k+1)} - \overline{x}_j) > \sum_{i:\underline{x}_i \ge x_{(k+1)}} m_i \cdot (\underline{x}_i - x_{(k+1)}). \tag{4}$$

  At each iteration of this bisection, we have an interval $[k^-, k^+]$ that is guaranteed to contain k. In the beginning, $k^- = 1$ and $k^+ = q$. At each stage, we compute the midpoint $k_{\text{mid}} = \lfloor (k^- + k^+)/2 \rfloor$, and check both inequalities (3) and (4) for $k = k_{\text{mid}}$. Then:

  - If both inequalities (3) and (4) hold for his k, this means that we have found the desired k.
  - If (3) holds but (4) does not hold, this means that the desired value k is larger than k_{mid}, so we keep k^+ and replace k^- with $k_{\text{mid}} + 1$.

– If (4) holds but (3) does not hold, this means that the desired value k is smaller than $k_{\rm mid}$, so we keep k^- and replace k^+ with $k_{\rm mid}-1$.

- Once k is found, we compute

$$S_k \stackrel{\rm def}{=} \sum_{i:\underline{x}_i \ge x_{(k+1)}} m_i \cdot \underline{x}_i + \sum_{j:\overline{x}_j \le x_{(k)}} m_j \cdot \overline{x}_j, \tag{5}$$

and

$$\Sigma_k \stackrel{\rm def}{=} \sum_{i:\underline{x}_i \ge x_{(k+1)}} m_i + \sum_{j:\overline{x}_j \le x_{(k)}} m_j.$$

If $\Sigma_k = 0$, we take $\underline{V} = 0$; otherwise, we compute $r_k = S_k/\Sigma_k$, and then

$$\underline{V} = \sum_{j:\overline{x}_j \le x_{(k)}} m_i \cdot (\overline{x}_j - r_k)^2 + \sum_{i:\underline{x}_i \ge x_{(k+1)}} m_j \cdot (\underline{x}_i - r_k)^2.$$

Comment. In principle, it is possible that for all the values i, we have $\underline{x}_i < x_{(k+1)}$ and $x_{(k)} < \overline{x}_i$. In this case, Σ_k is the sum of an empty number of terms, i.e., by a usual definition of such a sum, $\Sigma_k = 0$. In this case, $\underline{V}$ is also the sum of an empty set of terms, i.e., 0.

Comment. For the case when $m_1 = \ldots = m_n = 1/n$, this DS-related algorithm coincides with the algorithm for computing $\underline{V}$ under interval uncertainty; see, e.g., [6]. The explanation for this coincidence is given in the proof of the algorithm's correctness.

The algorithm $\overline{\mathcal{V}}$ for computing $\overline{V}$ is as follows:

- First, we sort all n midpoints $\widetilde{x}_i = \dfrac{1}{2} \cdot (\underline{x}_i + \overline{x}_i)$ into a non-decreasing sequence. After this sorting, we can assume that the intervals $\mathbf{x}_i$ are sorted in such a way that $\widetilde{x}_1 \le \widetilde{x}_2 \le \ldots \le \widetilde{x}_n$. We take $\widetilde{x}_{n+1} = +\infty$.

  We say that k is *proper* if $\widetilde{x}_k > \widetilde{x}_{k-1}$ or $k = 1$.

  For each k, we denote by $l(k)$ the largest value for which $\widetilde{x}_l = \widetilde{x}_k$, and by $s(k)$, the smallest value for which $\widetilde{x}_s = \widetilde{x}_k$. (Hence, the value $s(k)$ is always proper.)

- Second, we use bisection to find the value k ($1 \le k \le n$) for which the following two inequalities hold:

$$\sum_{j=k}^{n} m_j \cdot (\overline{x}_j - \widetilde{x}_k) < \sum_{i=1}^{k-1} m_i \cdot (\widetilde{x}_k - \underline{x}_i); \tag{6}$$

$$\sum_{j=k}^{n} m_j \cdot (\overline{x}_j - \widetilde{x}_{k-1}) \ge \sum_{i=1}^{k-1} m_i \cdot (\widetilde{x}_{k-1} - \underline{x}_i). \tag{7}$$

At each iteration of this bisection, we have an interval $[k^-, k^+]$ that is guaranteed to contain k. In the beginning, $k^- = 1$ and $k^+ = n + 1$. At each stage, we compute the midpoint $k_{\rm mid} = \lfloor (k^- + k^+)/2 \rfloor$, and check both inequalities (6) and (7) for $k = k_{\rm mid}$. Then:

- If both inequalities (6) and (7) hold for this $k_{\rm mid}$, this means that we have found the desired k.
- If, for $k_{\rm mid}$, (6) holds but (7) does not hold, this means that the desired value k is smaller than $k_{\rm mid}$, so we keep k^- and replace k^+ with $k_{\rm mid} - 1$.
- If, for $k_{\rm mid}$, (7) holds but (6) does not hold, this means that the desired value k is larger than $k_{\rm mid}$, so we keep k^+ and replace k^- with $k_{\rm mid} + 1$.

Once k is found, we compute

$$E \stackrel{\rm def}{=} \sum_{i=1}^{k-1} m_i \cdot \underline{x}_i + \sum_{j=k}^{n} m_j \cdot \overline{x}_j,$$

and then

$$V = \sum_{i=1}^{k-1} m_i \cdot (\underline{x}_i - E)^2 + \sum_{j=k}^{n} m_j \cdot (\overline{x}_j - E)^2.$$

- Third, we use bisection to find the proper value k ($1 \le k \le n$) for which the following two inequalities hold:

$$\sum_{j=l(k)+1}^{n} m_j \cdot (\overline{x}_j - \widetilde{x}_k) \le \sum_{i=1}^{l(k)} m_i \cdot (\widetilde{x}_k - \underline{x}_i); \tag{8}$$

$$\sum_{j=k}^{n} m_j \cdot (\overline{x}_j - \widetilde{x}_k) \ge \sum_{i=1}^{k-1} m_i \cdot (\widetilde{x}_k - \underline{x}_i). \tag{9}$$

At each iteration of this bisection, we have an interval $[k^-, k^+]$ that is guaranteed to contain k. In the beginning, $k^- = 1$ and $k^+ = n$. At each stage, we compute the proper index $k_{\rm mid} = s(\lfloor (k^- + k^+)/2 \rfloor)$ corresponding to the midpoint, and check both inequalities (8) and (9) for the proper value $k = k_{\rm mid}$. Then:

- If both inequalities (8) and (9) hold for this $k_{\rm mid}$, this means that we have found the desired k.
- If, for $k_{\rm mid}$, (8) holds but (9) does not hold, this means that the desired value k is smaller than $k_{\rm mid}$, so we keep k^- and replace k^+ with $k_{\rm mid} - 1$.

– If, for $k_{\rm mid}$, (9) holds but (8) does not hold, this means that the desired value k is larger than $k_{\rm mid}$, so we keep k^+ and replace k^- with $l(k_{\rm mid})+1$.

Once k is found, we compute

$$V=\sum_{i=1}^{k-1}(\underline{x}_i-\widetilde{x}_k)^2+\sum_{j=k}^{n}(\overline{x}_j-\widetilde{x}_k)^2.$$

- Finally, as $\overline{V}$, we take the largest of the two values of V obtained on the second and on the third steps.

Comment. Since, as we have mentioned, even when $m_1=\ldots=m_n=1/n$, the DS-related bound $\overline{V}$ may be different from the corresponding interval bound, this algorithm is different from the algorithms for computing $\overline{V}$ under interval uncertainty.

Proposition 3. *The algorithms $\underline{\mathcal{V}}$ and $\overline{\mathcal{V}}$ always compute the endpoints of the range $V(K)=[\underline{V},\overline{V}]$ in time $O(n\cdot\log(n))$.*

Comment: how good are these algorithms? Since even simple sorting requires at least $O(n\cdot\log(n))$ steps (see, e.g., [1]), algorithms like this, that compute a bound of a statistical interval characteristic in $O(n\cdot\log(n))$ steps, can be considered a "golden standard" for such algorithms.

Proof of Proposition 3. This proof is similar to the proofs from [5, 7, 8, 10] in its use of convexity.

It is known that the variance $V(\rho)$ is an example of a *convex* statistical characteristic, i.e., a characteristic for which $C(\alpha\cdot\rho+(1-\alpha)\cdot\rho')\le\alpha\cdot C(\rho)+(1-\alpha)\cdot C(\rho')$ for every two distributions ρ and ρ' and for every $\alpha\in(0,1)$. For convex characteristics, the following is true:

Lemma. *Let C be a convex characteristic, and let*

$$K=\langle(\mathbf{x}_1,\ldots,\mathbf{x}_n),(m_1,\ldots,m_n)\rangle$$

be a Dempster-Shafer knowledge base. Then, among all probability distributions ρ which are consistent with K (i.e., all distributions of type (2)):

- *The smallest possible value $\underline{C}$ is attained when for each i from 1 to n, we use a 1-point distribution in which x_i is equal to some value with probability 1.*
- *The largest possible value $\overline{C}$ is attained when for each i from 1 to n, we use a 2-point distribution for x_i, in which x_i can only attain endpoint values $\underline{x}_i$ and $\overline{x}_i$.*

Proof of the Lemma. For simplicity, we will consider *discrete* distributions which are located on finitely many points. For example, the distribution ρ_1 is located at points $x_1^{(1)}, \ldots, x_1^{(N_1)}$ with probabilities, correspondingly, $p_1^{(1)}, \ldots, p_1^{(N_1)}$; similarly, for every i, the distribution ρ_i is located at points $x_i^{(1)}, \ldots, x_i^{(N_i)}$ with probabilities, correspondingly, $p_i^{(1)}, \ldots, p_i^{(N_i)}$;

With respect to computing expectations like $V(\rho)$, every continuous distribution can be approximated, with arbitrary accuracy, by discrete ones. So, the result holds for all possible probability distributions. (Alternative, we can consider continuous distributions and integrals instead of sums; the proof will remain largely the same.)

Let us replace the two values $x_i^{(j)}$ and $x_i^{(j')}$ whose probabilities are, correspondingly, $m_i \cdot p_i^{(j)}$ and $m_i \cdot p_i^{(j')}$, with their convex combination

$$x_i \stackrel{\text{def}}{=} \alpha \cdot x_i^{(j)} + (1-\alpha) \cdot x_i^{(j')}$$

(where $\alpha \stackrel{\text{def}}{=} p_i^{(j)}/(p_i^{(j)}+p_i^{(j')})$) whose probability is $m_i \cdot p_i$, where $p_i \stackrel{\text{def}}{=} p_i^{(j)}+p_i^{(j')}$. Since the characteristic C is convex, the resulting value C can only decrease.

Since both values $x_i^{(j)}$ and $x_i^{(j')}$ were in the interval $[\underline{x}_j, \overline{x}_j]$, their convex combination x_i is also in this same interval.

Thus, if the minimum of C is attained for some distribution ρ_i which is located at N_i different points $x_i^{(j)}$, then we can replace two different values by their convex combination and attain the same minimum by using a distribution located at $N_i - 1$ different points. We can repeat this reduction again and again until we reach a distribution located at a single point x_i.

So, the minimum of $C(\rho)$ is indeed attained when each distribution ρ_i ($1 \le i \le n$) is located at a single point x_i with probability 1.

Similarly, if, in the distribution ρ_i, we replace a point $x_i^{(j)}$ whose probability is $p_i^{(j)}$ by two points $\underline{x}_i$ and $\overline{x}_i$ with probabilities $p_i^{(j)} \cdot \alpha$ and $p_i^{(j)} \cdot (1-\alpha)$, where $\alpha \stackrel{\text{def}}{=} \dfrac{\overline{x}_i - x_i^{(j)}}{\overline{x}_i - \underline{x}_i}$, then $x_i^{(j)} = \alpha \cdot \underline{x}_i + (1-\alpha) \cdot \overline{x}_i$. Therefore, due to convexity of C, this replacement can only increase the corresponding value of $C(\rho)$.

Thus, if the maximum of C is attained for some distribution ρ_i in which I_i internal points $x_i^{(j)} \in (\underline{x}_i, \overline{x}_i)$ have non-zero probabilities, then, by replacing each internal point $x_i^{(j)}$ with a probabilistic combination of endpoints $\underline{x}_i$ and $\overline{x}_i$, we get a new distribution that attains the same maximum and in which only $I_i - 1$ internal points have non-zero probability. We can repeat this reduction again and again until we reach a distribution located only at the endpoints $\underline{x}_i$ and $\overline{x}_i$. This is exactly the distribution described in the Lemma; thus, the Lemma is proven.

Proof of the result about $\underline{V}$. Due to the Lemma, the smallest possible value $\underline{V}$ of the variance V is attained when for each i, the distribution ρ_i is located at some point $x_i \in [\underline{x}_i, \overline{x}_i]$ with probability 1. One can check that in this case,

the variance takes the form $V = \sum_{i=1}^{n} m_i \cdot x_i^2 - E^2$, where $E = \sum_{i=1}^{n} m_i \cdot x_i$. So, to find the value $\underline{V}$, we must find the values $x_i \in \mathbf{x}_i$ for which this expression is the smallest possible.

One can see that for the case when $m_1 = \ldots = m_n = 1/n$, the above expression for V coincides with the standard expression for the population variance. For this case, an algorithm for computing the smallest possible value of V is given in [6]. In view of this fact, to solve our minimization problem, we will generalize the algorithm from [6] to this more general case of possible different masses m_i.

Similarly to [6], let us start with simple calculus. Let $f(x_1, \ldots, x_n)$ be a differentiable function on a box $B \stackrel{\text{def}}{=} \mathbf{x}_1 \times \ldots \times \mathbf{x}_n$, and let $x^- = (x_1^-, \ldots, x_n^-) \in B$ be a point at which f attains its smallest value on this box.

Then, for every i, the function $f_i(x_i) \stackrel{\text{def}}{=} f(x_1^-, \ldots, x_{i-1}^-, x_i, x_{i+1}^-, \ldots, x_n^-)$ also attains its minimum on the interval $[\underline{x}_i, \overline{x}_i]$ at the point $x_i = x_i^-$.

According to the basic calculus, this minimum is either attained in the interior of the interval, in which case $df_i/dx_i = 0$ for $x_i = x_i^-$, or the minimum is attained at one of the endpoints of the interval $[\underline{x}_i, \overline{x}_i]$. If the minimum is attained at the left endpoint $\underline{x}_i$, then the function f_i cannot be decreasing at this point, so $df_i/dx_i \geq 0$. Similarly, if the minimum is attained at the right endpoint $\overline{x}_i$, then $df_i/dx_i \leq 0$.

By definition of the function $f_i(x_i)$, the value of the derivative df_i/dx_i for $x_i = x_i^-$ is equal to the value of the partial derivative $\partial f/\partial x_i$ at the point x^-. Thus, for each i, we have one of the following three cases:

- either $\underline{x}_i < x_i^- < \overline{x}_i$ and $\partial f/\partial x_i = 0$;
- or $x_i^- = \underline{x}_i$ and $\partial f/\partial x_i \geq 0$;
- or $x_i^- = \overline{x}_i$ and $\partial f/\partial x_i \leq 0$.

For $f = V$, as one can easily see, $\partial V/\partial x_i = 2m_i \cdot (x_i - E)$, so the sign of this derivative is the same as the sign of the difference $x_i - E$. Therefore, for the point x^- at which the variance V attains its minimum, we have one of the following three situations:

- either $\underline{x}_i < x_i^- < \overline{x}_i$ and $x_i^- = E$;
- or $x_i^- = \underline{x}_i$ and $x_i^- \geq E$;
- or $x_i^- = \overline{x}_i$ and $x_i^- \leq E$.

In the first case, $\underline{x}_i < E < \overline{x}_i$; in the second case, $E \leq \underline{x}_i$, and in the third case, $\overline{x}_i \leq E$.

Let us show that if we know where E is in comparison to the endpoints of all the intervals, i.e., to which "zone" $[x_{(k)}, x_{(k+1)}]$ the value E belongs, we can uniquely determine the values x_i^- for all i.

Indeed, when $x_{(k+1)} \leq \underline{x}_i$, this means that $E \leq x_{(k+1)} \leq \underline{x}_i$, so $E \leq \underline{x}_i$. Thus, we cannot have the first case (in which $E > \underline{x}_i$), so we must have either

the second or the third cases, i.e., we must have $x_i = \underline{x}_i$ or $x_i = \overline{x}_i$. If the interval $\mathbf{x}_i$ is degenerate, then both cases lead to the same result. If the interval is non-degenerate, then we cannot have the third case – in which $\underline{x}_i < \overline{x}_i \le E$ hence $\underline{x}_i < E$ – and thus, we must have the second case, i.e., $x_i^- = \overline{x}_i$. Thus, $x_{(k+1)} \le \underline{x}_i$ implies that $x_i^- = \underline{x}_i$.

Similarly, $x_{(k)} \ge \overline{x}_i$ implies that $x_i^- = \overline{x}_i$, and in all other cases, we have $x_i^- = E$.

All that remains is to find the appropriate k. Once k is fixed, we can find the values x_i^- in linear time, and then compute the corresponding value V in linear time. The only condition on k is that the average of the corresponding values x_i^- should be within the corresponding zone $[x_{(k)}, x_{(k+1)})$.

In principle, we can find k by exhaustive (linear) search. Since there are $2n$ possible small intervals, we must therefore repeat $O(n)$ computations $2n$ times, which takes $2n \cdot O(n) = O(n^2)$ time. Together with the original sorting – that takes $O(n \cdot \log(n))$ time – we thus get a quadratic time algorithm, since

$$O(n^2) + O(n \cdot \log(n)) = O(n^2).$$

Let us now show that we can find k faster. We want to satisfy the conditions $x^{(k)} \le E$ and $E < x^{(k+1)}$. The value E is the weighted average of all the values x_i^-, i.e., we have

$$E = S_k + (1 - \Sigma_k) \cdot E, \tag{10}$$

where S_k is defined by the formula (5) and Σ_k is defined in the description of the algorithm $\underline{\mathcal{V}}$. By moving all the terms proportional to E to the left-hand side of (10), we conclude that $\Sigma_k \cdot E = S_k$, i.e., that $E = S_k/\Sigma_k$ ($= r_k$; the case when $\Sigma_k = 0$ is handled later in this proof). The first desired inequality $x^{(k)} \le E$ thus takes the form $S_k/\Sigma_k \le x^{(k)}$, i.e., equivalently, $\Sigma_k \cdot x^{(k)} \le S_k$, i.e.,

$$\left(\sum_{i:\underline{x}_i \ge x_{(k+1)}} m_i + \sum_{j:\overline{x}_j \le x_{(k)}} m_j \right) \cdot x_{(k)} \le \sum_{i:\underline{x}_i \ge x_{(k+1)}} \underline{x}_i + \sum_{j:\overline{x}_j \le x_{(k)}} \overline{x}_j. \tag{11}$$

If we subtract $m_i \cdot x_{(k)}$ (or, correspondingly, $m_j \cdot x_{(k)}$) from each term in the right-hand side and move terms proportional to $\overline{x}_j - x_{(k)}$ is to the left-hand side of the inequality, we get the desired inequality (3).

When k increases, the left-hand side of the inequality (3) increases – because each term increases and new terms may appear. Similarly, the right-hand side of this inequality decreases with k. Thus, if this inequality holds for k, it should also hold for all smaller values, i.e., for $k-1$, $k-2$, etc.

Similarly, the second desired inequality $E < x^{(k+1)}$ takes the equivalent form (4). When k increases, the left-hand side of this inequality increases, while the right-hand side decreases. Thus, if this inequality is true for k, it is also true for $k+1$, $k+2$, ...

If both inequalities (3) and (4) are true for two different values $k < k'$, then they should both be true for all the values intermediate between k and k', i.e.,

for $k+1, k+2, \ldots, k'-1$. Let us show that both inequalities cannot be true for k and for $k+1$. Indeed, if the inequality (3) is true for $k+1$, this means that

$$\sum_{j:\overline{x}_j \le x_{(k+1)}} m_j \cdot (x_{(k+1)} - \overline{x}_j) \le \sum_{i:\underline{x}_i \ge x_{(k+2)}} m_i \cdot (\underline{x}_i - x_{(k+1)}). \qquad (12)$$

However, the left-hand side of this inequality is not smaller than the left-hand side of (4), while the right-hand side of this inequality is not larger than the right-hand side of (4). Thus, (12) is inconsistent with (4). This inconsistency proves that there is only one k for which both inequalities are true, and this k can be found by the bisection method as described in the above algorithm $\underline{\mathcal{V}}$.

How long does this algorithm take? In the beginning, we only know that k belongs to the interval $[1, 2n]$ of width $O(n)$. At each stage of the bisection step, we divide the interval (containing k) in half. After I iterations, we decrease the width of this interval by a factor of 2^I. Thus, to find the exact value of k, we must have I for which $O(n)/2^I = 1$, i.e., we need $I = O(\log(n))$ iterations. On each iteration, we need $O(n)$ steps, so we need a total of $O(n \cdot \log(n))$ steps. With $O(n \cdot \log(n))$ steps for sorting, and $O(n)$ for computing the variance, we get a $O(n \cdot \log(n))$ algorithm. The statement about the algorithm $\underline{\mathcal{V}}$ is proven.

Comment. In the above text, we considered the case when $\Sigma_k \neq 0$. In a comment after the description of the algorithm for computing $\underline{V}$, we have mentioned that it is possible to have $\Sigma_k = 0$, i.e., it is possible that for all the values i, we have $\underline{x}_i < x_{(k+1)}$ and $x_{(k)} < \overline{x}_i$.

In this case, since the values $x_{(k)}$ are sorted endpoints $\underline{x}_i$ and $\overline{x}_i$, from the fact that $\underline{x}_i < x_{(k+1)}$, we conclude that $\underline{x}_i \le x_{(k)}$ – since $x_{(k)}$ is the largest of the endpoints which are smaller than $x_{(k+1)}$.

Similarly, $x_{(k)} < \overline{x}_i$ implies that $x_{(k+1)} \le \overline{x}_i$. Therefore, in this case, $\underline{x}_i \le x_{(k)} \le x_{(k+1)} \le \overline{x}_i$ for all i. Hence, all the intervals $\mathbf{x}_i$ contain the value $x_{(k)}$. If on each interval $\mathbf{x}_i$, we take a distribution that is located at $x_{(k)}$ with probability 1, we get the resulting 1-point distribution for which $V = 0$. Thus, in this case, indeed $\underline{V} = 0$ (in accordance with the above algorithm).

Proof of the result about $\overline{V}$. Due to the Lemma, the largest possible value $\overline{V}$ of the variance V is attained when for each i, the distribution ρ_i is located at two points: $\overline{x}_i$ and $\underline{x}_i$. Let p_i denote the probability of $\overline{x}_i$; then the probability of $\underline{x}_i$ is equal to $1 - p_i$. One can check that in this case, the variance takes the form

$$V = \sum_{i=1}^{n} m_i \cdot (p_i \cdot \overline{x}_i^2 + (1 - p_i) \cdot \underline{x}_i^2) - E^2,$$

where

$$E = \sum_{i=1}^{n} m_i \cdot (p_i \cdot \overline{x}_i + (1 - p_i) \cdot \underline{x}_i).$$

So, to find the value $\overline{V}$, we must find the values $p_i \in [0, 1]$ for which this expression V is the largest possible.

Let us apply the calculus-based analysis to the above problem of maximizing the expression V as a function of n variables $p_1, \ldots, p_n$. Here,

$$\frac{\partial V}{\partial p_i} = m_i \cdot (\overline{x}_i^2 - \underline{x}_i^2) - 2 \cdot E \cdot (\overline{x}_i - \underline{x}_i) =$$

$$2m_i \cdot (\overline{x}_i - \underline{x}_i) \cdot \left(\frac{\underline{x}_i + \overline{x}_i}{2} - E \right) = 2m_i \cdot (\overline{x}_i - \underline{x}_i) \cdot (\widetilde{x}_i - E),$$

where $\widetilde{x}_i$ is the midpoint of the interval $\mathbf{x}_i$. So, the sign of this derivative coincides with the sign of the difference $\widetilde{x}_i - E$. Thus, similarly to the case of $\underline{V}$, from the fact that V attains maximum, we conclude that for every i, we have three possible situations:

- either $0 < p_i < 1$ and $\widetilde{x}_i = E$;
- or $p_i = 0$ and $\widetilde{x}_i \le E$;
- or $p_i = 1$ and $\widetilde{x}_i \ge E$.

Let us show that if we know where E is in comparison to the midpoints $\widetilde{x}_i$ of all the intervals, then we can uniquely determine almost all the values p_i – except a few with the same $\widetilde{x}_i$.

Indeed, when $\widetilde{x}_i > E$, then we cannot have neither the first case (in which $E = \underline{x}_i$) nor the second case, so we must the third case $p_i = 1$, i.e., we must have $x_i = \overline{x}_i$ with probability 1.

Similarly, when $\widetilde{x}_i < E$, then we have $p_i = 0$, i.e., we have $x_i = \underline{x}_i$ with probability 1.

When $\widetilde{x}_i = E$, then we cannot say anything about p_i: all we know is that we have $\overline{x}_i$ with some probability p_i and $\underline{x}_i$ with the probability $1 - p_i$.

In our algorithm, we have sorted the intervals in such a way that their midpoints form an increasing sequence. So, we can assume that the values $\widetilde{x}_i$ are already sorted. In principle, there are two possible cases:

- the mean value E corresponding to the optimal distribution is different from all the values $\widetilde{x}_i$, and
- the mean value E corresponding to the optimal distribution coincides with one of the values $\widetilde{x}_i$.

Let us show that both cases are indeed possible:

- If we have two intervals $[-5, -4]$ and $[4, 5]$ with probability $1/2$ each, then the mean value E must be within the interval $[(-5 + 4)/2, (5 - 4)/2] = [-0.5, 0.5]$ and therefore, cannot coincide with any of the midpoints -4.5 and 4.5.
- On the other hand, in the above-cited example where we have three intervals $[0, 1]$ with probability $1/3$ each, we must have $E = \widetilde{x}_i$ for some i, because otherwise all three distributions ρ_i would be concentrated on one of the endpoints, and we already know that this way, we cannot attain the maximum of $V(K)$.

Let us analyze these two cases one by one.

In the first case, let k denote the smallest integer for which $\widetilde{x}_k > E$. Then, according to the above description, we have $x_i = \underline{x}_i$ for $i < k$ and $x_j = \overline{x}_j$ for $j \geq k$, hence $E = \sum_{i=1}^{k-1} m_i \cdot \underline{x}_i + \sum_{j=k}^{n} m_j \cdot \overline{x}_j$. Our selection of k means that $\widetilde{x}_{k-1} \leq E < \widetilde{x}_k$. Substituting the expression for E into this double inequality, we get the inequalities described in the algorithm.

Similar to the proof of correctness for the algorithm $\underline{\mathcal{V}}$, we can conclude that there is only one such k, and that the corresponding value k can indeed be found by the bisection described in the algorithm.

In the second case, let k be the first value for which $E = \widetilde{x}_k$. By definition of k, we must have $\widetilde{x}_k > \widetilde{x}_{k-1}$, so this k is a proper value. Let us recall that for each k, by $l(k)$ we denoted the largest index for which $\widetilde{x}_{l(k)} = \widetilde{x}_k$. Then, we have

$$E = \widetilde{x}_k = \sum_{i=1}^{k-1} m_i \cdot \underline{x}_i + \sum_{i=k}^{l(k)} m_i \cdot E_i + \sum_{j=l(k)+1}^{n} m_j \cdot \overline{x}_j,$$

where by E_i, we denoted the mean of $\rho_i(x)$. Since $E_i \in [\underline{x}_i, \overline{x}_i]$, we can find the interval of possible values of the right-hand side of this expression – namely, to get the lower bound, we replace E_i with $\underline{x}_i$, and to get the upper bound, we replace E_i with the upper bound $\overline{x}_i$. Thus, we conclude that the actual value $\widetilde{x}_k$ must be between the endpoints of this interval:

$$\sum_{i=1}^{l(k)} m_i \cdot \underline{x}_i + \sum_{j=l(k)+1}^{n} m_j \cdot \overline{x}_j \leq \widetilde{x}_k \leq \sum_{i=1}^{k-1} m_i \cdot \underline{x}_i + \sum_{j=k}^{n} m_j \cdot \overline{x}_j.$$

Similarly to the proof for $\underline{V}$, we can now conclude that Part 3 of the algorithm describes how to find the corresponding value k.

We will just mention that when $\widetilde{x}_k = E$, then $(\underline{x}_i - E)^2 = (\overline{x}_i - E)^2$, hence, no matter what p_i is, the corresponding two terms

$$m_i \cdot p_i \cdot (\overline{x}_i - E)^2 + m_i \cdot (1 - p_i) \cdot (\underline{x}_i - E)^2$$

in the expression for the variance always add up to the same value $m_i \cdot (\overline{x}_i - E)^2$.

The proposition is proven.

Comment. Similar algorithms can be described not only for the variance, but also for the characteristic $C = E + k_0 \cdot \sigma$ (where $\sigma = \sqrt{V}$ and k_0 is a fixed number), a characteristic which is useful in describing confidence intervals and outliers; see, e.g., [9, 13].

For C, the Lemma is still true: indeed, replacing two points with their mean decreases σ and leaves E intact, hence decreases C as well. Thus, in this case, the minimum of C is also attained for 1-point distributions; so we can use a natural generalization of interval algorithms from [9] to describe this more general case as well.

For C, the maximum is also attained for two-point distributions. Differentiating the resulting expression for C w.r.t. p_i, we conclude that the sign of the derivative coincides with the sign of the difference $\widetilde{x}_i - E'$ for some linear combination E' of E and σ. So, once we know where E' is in relation to the midpoints, we can make a similar conclusion about the maximizing distributions ρ_i – the only difference is that now the formulas expressing E' in terms of the selected values x_i are more complex.

5 Conclusions

In many real-life situations, we only have partial information about the actual probability distribution. For example, under Dempster-Shafer uncertainty, we only know the masses $m_1, \ldots, m_n$ assigned to different sets $S_1, \ldots, S_n$, but we do not know the distribution within each set S_i. Because of this uncertainty, there are many possible probability distributions consistent with our knowledge; different distributions have, in general, different values of standard statistical characteristics such as mean and variance. It is therefore desirable, given a Dempster-Shafer knowledge base, to compute the *ranges* of possible values of mean and of variance.

The existing algorithms for computing the range for the variance require $\approx 2^n$ computational steps, and therefore, cannot be used for large n. In this paper, we propose new efficient algorithms that work for large n as well.

It is worth mentioning that while for the Dempster-Shafer uncertainty, there exist efficient algorithms for computing the range of the variance, in a similar situation of interval uncertainty, the problem of computing the range for variance is NP-hard. Thus, with respect to computing the values (and ranges) of statistical characteristics, the case of Dempster-Shafer uncertainty is computationally simpler than the case of interval uncertainty.

Acknowledgments

This work was supported in part by NASA under cooperative agreement NCC5-209, by NSF grant EAR-0225670, by NIH grant 3T34GM008048-20S1, and by Army Research Lab grant DATM-05-02-C-0046

References

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

[2] R. E. Edwards, *Functional analysis: theory and applications*, Dover Publ., N.Y., 1995.

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

[4] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, Exact Bounds on Finite Populations of Interval Data, *Reliable Computing*, 2005, 11(3):207–233.

[5] S. Ferson, L. Ginzburg, V. Kreinovich, and J. Lopez, "Absolute Bounds on the Mean of Sum, Product, etc.: A Probabilistic Extension of Interval Arithmetic", *Extended Abstracts of the 2002 SIAM Workshop on Validated Computing*, Toronto, Canada, May 23–25, 2002, pp. 70–72.

[6] 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, editors, *Numerical Software with Result Verification*, Proceedings of the International Dagstuhl Seminar, Dagstuhl Castle, Germany, January 19–24, 2003, Springer Lectures Notes in Computer Science, 2004, Vol. 2991, pp. 274–305.

[7] V. Kreinovich, S. Ferson, and L. Ginzburg, Exact Upper Bound on the Mean of the Product of Many Random Variables With Known Expectations, *Reliable Computing*, 2003, 9(6):441–463.

[8] 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.

[9] V. Kreinovich, L. Longpré, P. Patangay, S. Ferson, and L. Ginzburg, Outlier Detection Under Interval Uncertainty: Algorithmic Solvability and Computational Complexity, Reliable Computing, 2005, 11(1):59–76.

[10] 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, June 30–July 2, 2004, Vol. 1, pp. 137–142.

[11] A. T. Langewisch and F. F. Choobineh, Mean and variance bounds and propagation of uncertainty for ill-specified random variables, *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 2004, 34(4):494–506.

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

[13] H. M. Wadsworth, Jr. (ed.), *Handbook of statistical methods for engineers and scientists*, McGraw-Hill Publishing Co., N.Y., 1990.

[14] R. R. Yager, J. Kacprzyk, and M. Pedrizzi (Eds.), *Advances in the Dempster-Shafer Theory of Evidence*, Wiley, N.Y., 1994