Which Value $\tilde{x}$ Best Represents a Sample $x_1, \ldots, x_n$: Utility-Based Approach Under Interval Uncertainty

Andrzej Pownuk and Vladik Kreinovich

Computational Science Program
University of Texas at El Paso
El Paso, TX 79968, USA
ampownuk@utep.edu, vladik@utep.edu

Abstract. In many practical situations, we have several estimates $x_1, \ldots, x_n$ of the same quantity $x$. In such situations, it is desirable to combine this information into a single estimate $\tilde{x}$. Often, the estimates $x_i$ come with interval uncertainty, i.e., instead of the exact values $x_i$, we only know the intervals $[\underline{x}_i, \overline{x}_i]$ containing these values. In this paper, we formalize the problem of finding the combined estimate $\tilde{x}$ as the problem of maximizing the corresponding utility, and we provide an efficient (quadratic-time) algorithm for computing the resulting estimate.

1 Which Value $\tilde{x}$ Best Represents a Sample $x_1, \ldots, x_n$: Case of Exact Estimates

Need to combine several estimates. In many practical situations, we have several estimates $x_1, \ldots, x_n$ of the same quantity $x$. In such situations, it is often desirable to combine this information into a single estimate $\tilde{x}$; see, e.g., [6].

Probabilistic case. If we know the probability distribution of the corresponding estimation errors $x_i - x$, then we can use known statistical techniques to find $\tilde{x}$, e.g., we can use the Maximum Likelihood Method; see, e.g., [8].

Need to go beyond the probabilistic case. In many cases, however, we do not have any information about the corresponding probability distribution [6]. How can we then find $\tilde{x}$?

Utility-based approach. According to the general decision theory, decisions of a rational person are equivalent to maximizing his/her utility value $u$; see, e.g., [1, 4, 5, 7]. Let us thus find the estimate $\tilde{x}$ for which the utility $u(\tilde{x})$ is the largest.

Our objective is to use a single value $\tilde{x}$ instead of all $n$ values $x_i$. For each $i$, the disutility $d = -u$ comes from the fact that if the actual estimate is $x_i$ and we use a different value $\tilde{x} \neq x_i$ instead, we are not doing an optimal thing. For example, if the optimal speed at which the car needs the least amount of fuel is $x_i$, and we instead run it at a speed $\tilde{x} \neq x_i$, we thus waste some fuel.
For each \( i \), the disutility \( d \) comes from the fact that the difference \( \tilde{x} - x_i \) is different from 0; there is no disutility if we use the actual value, so \( d = d(\tilde{x} - x_i) \) for an appropriate function \( d(y) \), where \( d(0) = 0 \) and \( d(y) > 0 \) for \( y \neq 0 \).

The estimates are usually reasonably accurate, so the difference \( x_i - \tilde{x} \) is small, and we can expand the function \( d(y) \) in Taylor series and keep only the first few terms in this expansion:

\[
d(y) = d_0 + d_1 \cdot y + d_2 \cdot y^2 + \ldots
\]

From \( d(0) = 0 \) we conclude that \( d_0 = 0 \). From \( d(y) > 0 \) for \( y \neq 0 \) we conclude that \( d_1 = 0 \) (else we would have \( d(y) < 0 \) for some small \( y \)) and \( d_2 > 0 \), so \( d(y) = d_2 \cdot y^2 = d_2 \cdot (\tilde{x} - x_i)^2 \).

The overall disutility \( d(\tilde{x}) \) of using \( \tilde{x} \) instead of each of the values \( x_1, \ldots, x_n \) can be computed as the sum of the corresponding disutilities

\[
d(\tilde{x}) = \sum_{i=1}^{n} d(\tilde{x} - x_i)^2 = d_2 \cdot \sum_{i=1}^{n} (\tilde{x} - x_i)^2.
\]

Maximizing utility \( u(\tilde{x}) \) equivalent to minimizing disutility.

**The resulting combined value.** Since \( d_2 > 0 \), minimizing the disutility function is equivalent to minimizing the re-scaled disutility function

\[
D(\tilde{x}) \doteq \frac{d(\tilde{x})}{d_2} = \sum_{i=1}^{n} (\tilde{x} - x_i)^2.
\]

Differentiating this expression with respect to \( \tilde{x} \) and equating the derivative to 0, we get

\[
\tilde{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i.
\]

This is the well-known sample mean.

2 Case of Interval Uncertainty: Formulation of the Problem

**Formulation of the practical problem.** In many practical situations, instead of the exact estimates \( x_i \), we only know the intervals \([\underline{x}_i, \overline{x}_i]\) that contain the unknown values \( x_i \). How do we select the value \( x \) in this case?

**Towards precise formulation of the problem.** For different values \( x_i \) from the corresponding intervals \([\underline{x}_i, \overline{x}_i]\), we get, in general, different values of utility

\[
U(\tilde{x}, x_1, \ldots, x_n) = -D(\tilde{x}, x_1, \ldots, x_n),
\]
Which Value $x$ Best Represents a Sample $x_1, \ldots, x_n$

where $D(\tilde{x}, x_1, \ldots, x_n) = \sum_{i=1}^{n} (\tilde{x} - x_i)^2$. Thus, all we know is that the actual (unknown) value of the utility belongs to the interval $[U(\tilde{x}), \overline{U}(\tilde{x})] = [-D(\tilde{x}), -\overline{D}(\tilde{x})]$, where

$$D(\tilde{x}) = \min D(\tilde{x}, x_1, \ldots, x_n),$$

$$\overline{D}(\tilde{x}) = \max D(\tilde{x}, x_1, \ldots, x_n),$$

and min and max are taken over all possible combinations of values $x_i \in [x_i, \overline{x}_i]$.

In such situations of interval uncertainty, decision making theory recommends using Hurwicz optimism-pessimism criterion [2–4], i.e., maximize the value

$$U(\tilde{x}) \overset{\text{def}}{=} \alpha \cdot U(\tilde{x}) + (1 - \alpha) \cdot \overline{U}(\tilde{x}),$$

where the parameter $\alpha \in [0, 1]$ describes the decision maker’s degree of optimism. For $U = -D$, this is equivalent to minimizing the expression

$$D(\tilde{x}) = -U(\tilde{x}) = \alpha \cdot D(\tilde{x}) + (1 - \alpha) \cdot \overline{D}(\tilde{x}).$$

What we do in this paper. In this paper, we describe an efficient algorithm for computing the value $\tilde{x}$ that minimizes the resulting objective function $D(\tilde{x})$.

3 Analysis of the Problem

Let us simplify the expressions for $D(\tilde{x})$, $\overline{D}(\tilde{x})$, and $D(\tilde{x})$. Each term $(\tilde{x} - x_i)^2$ in the sum $D(\tilde{x}, x_1, \ldots, x_n)$ depends only on its own variable $x_i$. Thus, with respect to $x_i$:

- the sum is the smallest when each of these terms is the smallest, and
- the sum is the largest when each term is the largest.

One can easily see that when $x_i$ is in the $[x_i, \overline{x}_i]$, the maximum of a term $(\tilde{x} - x_i)^2$ is always attained at one of the interval’s endpoints:

- at $x_i = \underline{x}_i$ when $\tilde{x} \geq \underline{x}_i \overset{\text{def}}{=} \frac{\underline{x}_i + \overline{x}_i}{2}$ and
- at $x_i = \overline{x}_i$ when $\tilde{x} < \underline{x}_i$.

Thus,

$$\overline{D}(\tilde{x}) = \sum_{i: \tilde{x} < \underline{x}_i} (\tilde{x} - \underline{x}_i)^2 + \sum_{i: \tilde{x} \geq \underline{x}_i} (\tilde{x} - \underline{x}_i)^2.$$

Similarly, the minimum of the term $(\tilde{x} - x_i)^2$ is attained:

- for $x_i = \tilde{x}$ when $\tilde{x} \in [\underline{x}_i, \overline{x}_i]$ (in this case, the minimum is 0);
- for $x_i = \underline{x}_i$ when $\tilde{x} < \underline{x}_i$; and
- for $x_i = \overline{x}_i$ when $\tilde{x} > \overline{x}_i$. 

Thus,
\[ D(\tilde{x}) = \sum_{i : x > \pi_i} (\tilde{x} - \pi_i)^2 + \sum_{i : x < \xi_i} (\tilde{x} - \xi_i)^2. \]
So, for \( D(\tilde{x}) = \alpha \cdot D(\tilde{x}) + (1 - \alpha) \cdot \overline{D}(\tilde{x}) \), we get
\[ D(\tilde{x}) = \alpha \cdot \sum_{i : x > \pi_i} (\tilde{x} - \pi_i)^2 + \alpha \cdot \sum_{i : x < \xi_i} (\tilde{x} - \xi_i)^2 + (1 - \alpha) \cdot \sum_{i : x > \pi_i} (\tilde{x} - \pi_i)^2 + (1 - \alpha) \cdot \sum_{i : x < \xi_i} (\tilde{x} - \xi_i)^2. \]

\[ (1) \]

Towards an algorithm. The presence or absence of different values in the above expression depends on the relation of \( \tilde{x} \) with respect to the values \( \pi_i, \pi_i, \) and \( \tilde{x} \). Thus, if we sort these \( 3n \) values into a sequence \( s_1 \leq s_2 \leq \ldots \leq s_{3n} \), then on each interval \([s_j, s_{j+1}]\), the function \( D(\tilde{x}) \) is simply a quadratic function of \( \tilde{x} \).

A quadratic function attains its minimum on an interval either at one of its endpoints or at a point when the derivative is equal to 0 (if this point is inside the given interval). Differentiating the above expression for \( D(\tilde{x}) \), equating the derivative to 0, dividing both sides by 0, and moving terms proportional not containing \( \tilde{x} \) to the right-hand side, we conclude that
\[ (\alpha \cdot \#\{i : \tilde{x} < \xi_i \text{ or } \tilde{x} > \pi_i\} + 1 - \alpha) \cdot \tilde{x} = \alpha \cdot \sum_{i : \tilde{x} > \pi_i} \pi_i + \alpha \cdot \sum_{i : \tilde{x} < \xi_i} \xi_i + (1 - \alpha) \cdot \sum_{i : \tilde{x} < \xi_i} \pi_i + (1 - \alpha) \cdot \sum_{i : \tilde{x} > \pi_i} \xi_i. \]

Since \( s_j \) is a listing of all thresholds values \( \pi_i, \pi_i, \) and \( \tilde{x} \), then for \( \tilde{x} \in (s_j, s_{j+1}) \), the inequality \( \tilde{x} < \xi_j \) is equivalent to \( s_{j+1} \leq \pi_j \). Similarly, the inequality \( \tilde{x} > \pi_j \) is equivalent to \( s_j \geq \pi_j \). In general, for values \( \tilde{x} \in (s_j, s_{j+1}) \), the above equation gets the form
\[ (\alpha \cdot \#\{i : \tilde{x} < \xi_i \text{ or } \tilde{x} > \pi_i\} + 1 - \alpha) \cdot \tilde{x} = \alpha \cdot \sum_{i : s_{j+1} \leq \pi_i} \pi_i + \alpha \cdot \sum_{i : s_j \leq \xi_i} \xi_i + (1 - \alpha) \cdot \sum_{i : s_{j+1} \leq \pi_i} \pi_i + (1 - \alpha) \cdot \sum_{i : s_j \geq \pi_i} \xi_i. \]

From this equation, we can easily find the desired expression for the value \( \tilde{x} \) at which the derivative is 0.

Thus, we arrive at the following algorithm.

4 Resulting Algorithm

First, for each interval \([\xi_i, \pi_i]\), we compute its midpoint \( \tilde{x}_i = \frac{\xi_i + \pi_i}{2} \). Then, we sort the \( 3n \) values \( \xi_i, \pi_i, \) and \( \tilde{x}_i \) into an increasing sequence \( s_1 \leq s_2 \leq \ldots \leq s_{3n} \).

To cover the whole real line, to these values, we add \( s_0 = -\infty \) and \( s_{3n+1} = +\infty \).
We compute the value of the objective function (1) on each of the endpoints $s_1, \ldots, s_{3n}$. Then, for each interval $(s_i, s_{j+1})$, we compute the value
\[
\tilde{x} = \frac{\alpha \cdot \sum_{i:s_i \leq x_i} x_i + \alpha \cdot \sum_{i:s_{j+1} \leq x_i} x_i + (1 - \alpha) \cdot \sum_{i:s_i \leq x_{j+1}} x_i + (1 - \alpha) \cdot \sum_{i:s_{j+1} \geq x_i} x_i}{\alpha \cdot \# \{ i : \tilde{x} < x_i \text{ or } \tilde{x} > x_i \} + 1 - \alpha}.
\]
If the resulting value $\tilde{x}$ is within the interval $(s_i, s_{j+1})$, we compute the value of the objective function (1) corresponding to this $\tilde{x}$.

What is the computational complexity of this algorithm. Sorting $3n = O(n)$ values $x_i$, $\pi_i$, and $\tilde{x}_i$ takes time $O(n \cdot \ln(n))$.

Computing each value $D(\tilde{x})$ of the objective function requires $O(n)$ computational steps. We compute $D(\tilde{x})$ for $3n$ endpoints and for $\leq 3n + 1$ values at which the derivative is 0 at each of the intervals $(s_j, s_{j+1})$ – for the total of $O(n)$ values.

Thus, overall, we need $O(n \cdot \ln(n)) + O(n) \cdot O(n) = O(n^2)$ computation steps. Hence, our algorithm runs in quadratic time.

Acknowledgments. This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, and by an award “UTEP and Prudential Actuarial Science Academy and Pipeline Initiative” from Prudential Foundation.

References