How to Detect Possible Additional Outliers: Case of Interval Uncertainty[*]

Hani Dbouk, Steffen Schön, Ingo Neumann
Institute of Geodesy
Leibniz University of Hannover
30167 Hannover, Germany
dbouk@ife.uni-hannover.de,schoen@ife.uni-hannover.de

neumann@gih.uni-hannover.de

Vladik Kreinovich[†]
Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
vladik@utep.edu

Abstract

In many practical situations, measurements are characterized by interval uncertainty – namely, based on each measurement result, the only information that we have about the actual value of the measured quantity is that this value belongs to some interval. If several such intervals – corresponding to measuring the same quantity – have an empty intersection, this means that at least one of the corresponding measurement results is an outlier, caused by a malfunction of the measuring instrument. From the purely mathematical viewpoint, if the intersection is non-empty, there is no reason to be suspicious. However, from the practical viewpoint, if the intersection is too narrow – i.e., almost empty – then we should also be suspicious, and mark this as an possible additional outlier case. In this paper, we describe a natural way to formalize this idea, and an algorithm for detecting such additional possible outliers.

Keywords: interval uncertainty, outliers, probabilistic approach
AMS subject classifications: 65G30, 65G40, 62P30

[*] Submitted: May 30, 2021; Revised: June 21, 2021; Accepted: June 22, 2021.

[†] This work was supported by the German Research Foundation (DFG) as a part of the Research Training Group i.c.sens (grant GRK2159) and by the Institute of Geodesy of the Leibniz University of Hannover. It was also supported in part by the US National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science) and HRD-1242122 (Cyber-ShARE Center of Excellence). This paper was written when V. Kreinovich was visiting Leibniz University of Hannover. The authors are grateful to the anonymous referees for valuable suggestions.

1 Formulation of the Problem

Need for several measurements of the same quantity. Most of the information about the putside world comes from measurements. Measurement are never 100% accurate: the measurement result $\widetilde{x}$ is, in general, different from the actual (unknown) value of the measured quantity. In other words, the measurement error $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$ is, in general, different from 0. Measurements are also not 100% reliable: sometimes measuring instruments malfunction.

A natural way to increase the accuracy and reliability of our information is to perform several measurements of the same quantity.

Which measurement results are outliers: an important problem. The fact that measurement are not 100% reliable means that sometimes measuring instruments malfunction – e.g., get stuck in the previously measured value. If we view such a measurement result as reflecting the true value of the measured quantity, we will get a false impression – and we may make bad decision based on this impression. For example, if the temperature in the chemical reactor starts rising above the optimal level, we need to cool it down to avoid it getting into an ineffective regime or even blowing up. However, if the temperature sensor is stuck in the previously measured (normal) value, we will not notice this potential dangerous development. Similarly, if a distance-measuring sensor in a self-driving vehicle gets stuck in the previous value of the distance from the vehicle to a nearby wall, this malfunction may lead to the vehicle hitting this wall.

In all such cases, it is desirable to decide whether all the measurement results are correct or whether some of them are suspicious – possibly outliers.

Comment. This problem is important not only for *direct* measurements, when the measurement values come directly from the corresponding sensors, but also for *indirect* measurements, when the estimate $\widetilde{x}$ for the quantity of interest x comes from applying an appropriate algorithm to one or several results of direct measurement.

Need to consider interval uncertainty. Detecting outliers is one of the main problem in measurements; see, e.g., [5, 6]. In metrology, this problem is usually analyzed in the probabilistic case – when we know the probability distribution of the measurement errors [5, 6]. However, in many practical situations, the only information that we have about the measurement error Δx is an upper bound Δ on its absolute value $|\Delta x|$: $|\Delta x| \le \Delta$. This upper bound is usually called the *accuracy* of this measurement.

In this case, once we know the measurement result $\widetilde{x}$, the only think we can conclude about the actual value x is that this value belongs to the interval

$$[\widetilde{x} - \Delta, \widetilde{x} + \Delta].$$

How can we detect outliers under such interval uncertainty?

Detecting definite outliers under interval uncertainty. After n measurements, we get n intervals describing the same quantity. If all measurements are correct – i.e., if none of them was an outlier – then the actual value x belongs to all these intervals and thus, belongs to their intersection. So, in this case, the intersection is non-empty. Thus, if the intersection of n intervals obtained from n measurements is empty, this means that at least one of the measurements is an outlier; see, e.g., [3, 7, 8].

Remaining problem: detecting possible additional outliers. At first glance, it may seem that if the intersection of all the intervals is non-empty, then we have no reason to conclude that one of the measurement results was an outlier. However, simple examples show that even when the intersection is non-empty, we may have a good reason to suspect that one of the measurements was an outlier.

Indeed, let us assume that we have performed two measurements with accuracy $\Delta = 1$, resulting in values $\widetilde{x} = 0$ and $\widetilde{x} = 2$, i.e., in intervals $[0 - 1, 0 + 1] = [-1, 1]$ and $[2-2, 2+1] = [1, 3]$. The intersection of these two intervals is non-empty: it consists of a single point $[-1, 1] \cap [1, 3] = \{1\}$. However, intuitively, something is not right here: we started with two not very accurate measurements, with accuracy ±1 comparable with the actual values, and we magically got a very accurate result? It is much more probable that one of these measurements is an outlier.

Similarly, if we got measurement results 0 and 1.9, with intervals $[-1, 1]$ and $[0.9, 2.9]$, the intersection of these two intervals is non-empty – it is equal to the interval $[0.9, 1]$. However, intuitively, this situation is suspicious: we started with measurements of low accuracy, and suddenly magically we got a 10 times more accurate estimate?

It is desirable to flag such suspicious cases, when the intersection is non-empty but still, we have good reasons to believe that one of the measurement results was an outlier.

What we do in this paper. In this paper, we provide a possible method for formally describing such suspicions and thus, to detecting such suspicious cases.

2 How to Formulate and Solve This Problem

Problem: reminder. We have several results $\widetilde{x}_1, \ldots, \widetilde{x}_n$ of measuring the same quantity x. We also know the accuracies $\Delta_1, \ldots, \Delta_n$ of these measurements. If the intersection of the corresponding intervals $[\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$ is empty, then clearly one of the measurement result was an outlier. In this paper, we consider situations when all the intervals have a non-empty intersection.

Idea. In general, the width w of the intersection cannot exceed the smallest of the widths $2\Delta_i$ of measurement-related intervals: the width w is either equal to this smallest width, or is smaller than the smallest of the original widths. We want to mark cases when the width w of the intersection is improbably small, much smaller than the smallest width of the corresponding intervals.

In general, we can have different intersection widths W with different probabilities. We want to dismiss the situations when, for the actually observed intersection width w, the probability $p(w) = \text{Prob}(W \le w)$ is smaller than a certain threshold p_0. This is a usual idea in applications of probability, where we routinely dismiss hypotheses whose probability is too small: e.g., smaller than 5%, or smaller than 1%, or smaller than 0.1% (see, e.g., [6]).

How to formalize this idea. How can we determine the desired probability in the case of interval uncertainty? The actual value x of the measured quantity can be any number from the interval $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$. There is no reason to assume that some values from this interval are more probable and some are less probable – it is therefore reasonable to assume that all the values from this interval are equally probable, i.e., that the actual value is uniformly distributed on this interval. This argument – known as *Laplace Indeterminacy Principle* – is widely used in applications of statistics; see, e.g., [4, 6].

It is also reasonable to assume that measurement errors of different measurements are independent. Indeed, if two of the measurement errors were strongly correlated, there would have been no advantage of performing the second measurement: this second measurement would simply repeat the previous result. Thus, we arrive at the following formal description of this problem.

A formal description of the problem. Let x denote the actual value of the quantity. Then, if we perform n measurements with accuracies $\Delta_1, \ldots, \Delta_n$, the measurement results are equal to $\widetilde{x}_i = x_i + \Delta x_i$, where each Δx_i is uniformly distributed on the interval $[-\Delta_i, \Delta_i]$. The intersection of these intervals has the form

$$\left[\max_{i=1,\ldots,n}(x + \Delta x_i - \Delta_i), \min_{i=1,\ldots,n}(x + \Delta x_i + \Delta_i)\right] =$$

$$\left[x + \max_{i=1,\ldots,n}(\Delta x_i - \Delta_i), x + \min_{i=1,\ldots,n}(\Delta x_i + \Delta_i)\right].$$

The width W of this intersection is thus equal to

$$W = \min_{i=1,\ldots,n}(\Delta x_i + \Delta_i) - \max_{i=1,\ldots,n}(\Delta x_i - \Delta_i). \tag{1}$$

One can see that this expression does not depend on x. So, the problem takes the following form:

- given n values $\Delta_1, \ldots, \Delta_n$ and the value w and p_0,
- check whether $\text{Prob}(W \le w) \le p_0$, where W is described by the expression (1) and $\Delta x_1, \ldots, \Delta x_n$ are independent random variables each of which is uniformly distributed on the corresponding interval $[-\Delta_i, \Delta_i]$.

General case: Monte-Carlo algorithm. Once we know the values Δ_i and w, we can compute the desired probability $\text{Prob}(W \le w)$ by the following direct simulation.

We select some large number K. Then, for $k = 1, \ldots, K$:

- first, we use the usual random number generators to simulate, for $i = 1, \ldots, n$, the values $\Delta x_i^{(k)}$ which are uniformly distributed on the interval $[-\Delta_i, \Delta_i]$;
- then, we compute the width $W^{(k)}$ of the intersection

  $$\bigcap_{i=1}^{n}\left[\Delta x_i^{(k)} - \Delta_i, \Delta x_i^{(k)} + \Delta_i\right],$$

  i.e., the value

  $$W^{(k)} = \min_{i=1,\ldots,n}\left(\Delta x_i^{(k)} + \Delta_i\right) - \max_{i=1,\ldots,n}\left(\Delta x_i^{(k)} - \Delta_i\right),$$

  and we check whether $W^{(k)} \le w$.

We then estimate the desired probability $\text{Prob}(W \le w)$ as the ratio $K_{\le}/K$, where $K_{\le}$ is the number of indices k for which we had $W^{(k)} \le w$. If this ratio is smaller than or equal to the selected threshold p_0, then we conclude that one of the measurement results was a possible additional outlier.

In the case of two measurements, we can use a faster algorithm. In the simplest case when we have $n = 2$ measurements, we can have an explicit expression for the desired probability. Namely, for any value $w < \min(2\Delta_1, 2\Delta_2)$, we have

$$\text{Prob}(W \le w) = \frac{w^2}{4\Delta_1 \cdot \Delta_2}. \tag{2}$$

Example. In particular, in the above example, when $\Delta_1 = \Delta_2 = 1$ and $w = 1 - 0.9 = 0.1$, this probability is equal to 0.0025 – which is indeed very small, less than 1%.

Proof of the formula (2). Since both variables Δx_i are independent and uniformly distributed on the corresponding intervals $[-\Delta_i, \Delta_i]$, the probability that $W \le w$ is equal to the ratio $A_{\le}/A$, where:

- $A_{\le}$ is the area of the set of all the pairs $(\Delta x_1, \Delta_2)$ for which the inequality $W \le w$ is satisfied, and
- A is the area $A = (2\Delta_1) \cdot (2\Delta_2) = 4\Delta_1 \cdot \Delta_2$ of the whole box

$$[-\Delta_1, \Delta_1] \times [-\Delta_2, \Delta_2].$$

For $n = 2$, the width W of the intersection has the form

$$W = \min(\Delta x_1 + \Delta_1, \Delta x_2 + \Delta_2) - \max(\Delta x_1 - \Delta_1, \Delta x_2 - \Delta_2).$$

To find an explicit expression for the width, we need to decide:

- which of the two minimized expressions from the expression for the width is the smallest, and
- which of the two maximized expressions is the largest.

Here:

- the condition $\Delta x_1 + \Delta_1 \le \Delta x_2 + \Delta_2$ is equivalent to $\Delta x_1 - \Delta x_2 \le \Delta_2 - \Delta_1$, and
- the condition $\Delta x_1 - \Delta_1 \le \Delta x_2 - \Delta_2$ is equivalent to $\Delta x_1 - \Delta x_2 \le \Delta_1 - \Delta_2$.

Thus, to find all possible cases, it is sufficient to compare the difference $\Delta x_1 - \Delta x_2$ with the values $\Delta_1 - \Delta_2$ and $\Delta_2 - \Delta_1$.

Without losing generality, we can assume that the second measurement was more accurate (or of the same accuracy), i.e. that $\Delta_2 \le \Delta_1$. In this case,

$$\Delta_2 - \Delta_1 \le 0 \le \Delta_1 - \Delta_2,$$

so we have $\Delta_2 - \Delta_1 \le \Delta_1 - \Delta_2$. Thus, there are three possible results of such comparisons:

1. we can have $\Delta_1 - \Delta_2 \le \Delta x_1 - \Delta x_2$;
2. we can have $\Delta_2 - \Delta_1 \le \Delta x_1 - \Delta x_2 \le \Delta_1 - \Delta_2$; and
3. we can have $\Delta x_1 - \Delta x_2 \le \Delta_2 - \Delta_1$.

In the second case, the width W of the intersection is equal to

$$(\Delta x_2 + \Delta_2) - (\Delta x_2 - \Delta_2) = 2\Delta_2.$$

This width is equal to the smaller of the two widths – i.e., the largest possible value of the intersection width, so it cannot be smaller than $w < \min(2\Delta_1, 2\Delta_2) = 2\Delta_2$. Thus, the condition $W \le w$ can only be satisfied in the first and third cases.

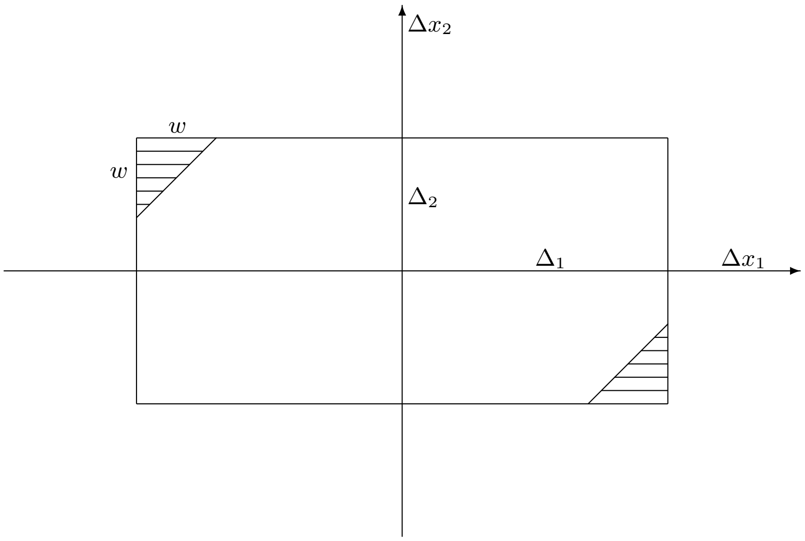
In the first case, the width is equal to

$$W = (\Delta x_2 + \Delta_2) - (\Delta x_1 - \Delta_1) = \Delta_1 + \Delta_2 - (\Delta x_1 - \Delta x_2).$$

Thus, the condition $W \le w$ is equivalent to

$$\Delta_1 - \Delta_2 - w \le \Delta x_1 - \Delta x_2. \tag{3}$$

The corresponding part of the box is shaded in the top left part of the following picture:



The set of all the pairs (Δ_1, Δ_2) for which this inequality is satisfied forms a right triangle with both sides equal to w, so its area is $w^2/2$. Similarly, the set corresponding to the third case – shaded in the bottom right part of the above picture – has area $w^2/2$. Thus:

- the overall area $A_\le$ is equal to $(w^2/2) + (w^2/2) = w^2$, and
- the desired probability $A_\le/A$ is indeed equal to the ratio $w^2/(4\Delta_1 \cdot \Delta_2)$.

The formula (2) is proven.

References

[1] H. Dbouk and S. Schön, "Guaranteed bounding zones for GNSS positioning by geometrical constraints", *Proceedings of the 11th Summer Workshop on Interval Methods SWIM'2018*, Rostock, Germany, July 25–27, 2018.

[2] H. Dbouk and S. Schön, "Reliability and integrity measures of GPS positioning via geometrical constraints", *Proceedings of the 2019 International Teachnical Meeting of the Institute of Navigation*, Reston, Virginia, January 28–31, 2019, pp. 730–743.

[3] L. Jaulin, "Probabilistic set-membership approach for robust regression", *Journal of Statistical Theory and Practice*, 2010, Vol. 4, pp. 155–167.

[4] E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

[5] S. G. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice*, Springer, New York, 2005.

[6] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.

[7] J. Sliwka, L. Jaulin, M. Ceberio, and V. Kreinovich, “Processing interval sensor data in the presence of outliers, with potential applications to localizing underwater robots”, *Proceedings of the 2011 IEEE International Conference on Systems, Man, and Cybernetics SMC’2011*, Anchorage, Alaska, October 9–12, 2011, pp. 2330–2337.

[8] A. Welte, L. Jaulin, M. Ceberio, and V. Kreinovich, “Robust data processing in the presence of uncertainty and outliers: case of localization problems”, *Proceedings of the IEEE Series of Symposia in Computational Intelligence SSCI’2016*, Athens, Greece, December 6–9, 2016.