What If There Are Too Many Outliers? *

Olga Kosheleva\textsuperscript{[0000–0003–2587–4209]} and Vladik Kreinovich\textsuperscript{[0000–0002–1244–1650]}

University of Texas at El Paso, El Paso, TX 79968, USA
{olgak, vladik}@utep.edu
https://www.cs.utep.edu/vladik/olgavita.html
https://www.cs.utep.edu/vladik/longvita.html

Abstract. Sometimes, a measuring instrument malfunctions, and we get a value which is very different from the actual value of the measured quantity. Such values are known as outliers. Usually, data processing considers situations when there are relatively few outliers. In such situations, we can delete most of them. However, there are situations when most results are outliers. In this case, we cannot produce a single value which is close to the actual value, but we can generate several values one of which is close. Of course, all the values produced by the measuring instrument(s) satisfy this property, but there are often too many of them, we would like to compress this set into a smaller-size one. In this paper, we prove that irrespective of the size of the original sample, we can always compress this sample into a compressed sample of fixed size.

Keywords: Outliers · Interval uncertainty · Ellipsoid uncertainty · Combinatorial convexity.

1 Formulation of the Problem

1.1 What are outliers and how they are treated now

Usually, measuring instruments work reliably, and produce a measurement result which is close to the actual value of the measured quantity. However, sometimes, measurement instruments malfunction, and the value they produce are drastically different from the actual value of the corresponding quantity. Such values are known as “outliers”.

When we process measurement results, it is important to delete as many outliers as possible. Indeed, if we take them at face value, we may get a biased

\* This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes), and by the AT&T Fellowship in Information Technology.

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).
impression of the situation and thus, based on this biased impression, we will make a wrong decision.

In some cases, outliers are easy to detect.

- If I step on a scale and get my weight as 10 kg, clearly something is wrong.
- If I measure my body temperature and the result is 30°C, this cannot be right.
- Similarly, if a patient has a clear fever, but the thermometer shows 36°C, something is wrong with this thermometer.

However, in many other situations, it is not as easy to detect outliers.

Usual methods for detecting outliers are based on the assumption that the majority of measurement results are correct. In this case, e.g., if we take the median of all the measurement results, this guarantees that this result will not be an outlier.

1.2 But what if there are too many outliers?

However, in some practical situations, it is the outliers that form the majority: the measuring instrument is malfunctioning most of the time, and the correct measurement results are in the minority.

The usual approach to such a situation is to ignore all the values and to try to improve the measuring instrument.

1.3 Formulation of the problem

If we made 1000 measurements and 60% of the results are outliers, still there are 400 correct measurement results. Clearly these results contain a lot of information about the studied system. It is therefore desirable to extract some information from these values.

How can we do it?

2 How to Extract Information: Idea

2.1 Why this extraction is not easy

When we have a small number of outliers, we can delete them and produce a value which is close to the actual value of the measured quantity. Unfortunately, in situations when the majority of results are outliers, this is not possible.

For example, suppose that:

- 1/3 of the measurement results are 0s,
- 1/3 of the measurement results are 1s,
- 1/3 of the measurement results are 2s,

and we know that no more than 2/3 of these results are outliers. In this case:
it could be that 0 is the actual value, and 1 and 2 are outliers;
• it could be that 1 is the actual value, and 0 and 2 are outliers; and
• it could be that 2 is the actual value, and 0 and 1 are outliers.

In such a situation, the only conclusion that we can make is that one of these three results 0, 1, and 2 is close to the actual value – and we do not know which one.

This list of possible values does not have to include all measurement results. For example, if we measure with accuracy 0.1, we know that no more than 2/3 of the results are outliers, and:

• 1/3 of the measurement results are 0s,
• 1/2 of the measurement results are 1s, and
• 1/6 of the measurement results are 2s,

then only 0 and 1 can be close to the actual value. Indeed, if 2 was the actual value, then we would have 5/6 outliers, which contradicts to our knowledge that the proportion of outliers does not exceed 2/3.

2.2 Main idea

As we have mentioned, in situations when there are too many outliers, we cannot select a single result which is close to the actual value of the measured quantity.

A natural idea is thus to extract a finite list of results so that one of them is close to the actual value – and ideally, we should make this list as small as possible.

3 Let Us Formulate This Problem in Precise Terms

3.1 Often, a sensor measures several quantities

A simple measuring instrument – such as a thermometer – measures only one quantity. However, many measuring instruments measure several quantities at the same time. For example, chemical sensors often measure concentrations of several substances, wind measurements usually involve measuring not only the wind’s speed but also its direction, sophisticated meteorological instruments measure humidity in addition to temperature, etc.

In principle, we can simply consider such a complex measuring instrument as a collection of several instruments measuring different quantities. However, if we do this, we will miss the fact that such an instrument usually malfunctions as a whole: if one of its values is an outlier, this means that this instrument malfunctioned, and we should not trust other values that it produced either. In other words, either all its values are correct, or all the values that this instrument produced are outliers. So, from the viewpoint of outliers, it make sense to consider it as a single measuring instrument producing several values.

In all such situations, as a result of the $j$-th measurement of the values of several ($d$) quantities, we get $d$ values $x_{j1}, \ldots, x_{jd}$. These values can be naturally represented by a point $x_j = (x_{j1}, \ldots, x_{jd})$ in the $d$-dimensional space.
3.2 How can we represent uncertainty

In the 1-D case, a usual information about the measurement error $\Delta x \overset{\text{def}}{=} x_j - x$, i.e., the difference between the measurement result $x_j$ and the (unknown) actual value $x$ of the desired quantity – is the lower bound $\Delta^- < 0$ and the upper bound $\Delta^+ > 0$ on its value: $\Delta^- \leq \Delta x_j \leq \Delta^+$; see, e.g., [12]. In this case, once we know the measurement result, we can conclude that the actual value $x$ is located somewhere in the set $x_j - [\Delta^-, \Delta^+]$, where, as usual, $x_j - U$ means the set of possible values $x_j - u$ when $u \in U$; see, e.g., [8–10, 12].

In the general case, when both $x_j$ and $x$ are $d$-dimensional, the measurement error $\Delta x_j = x_j - x$ is also $d$-dimensional. We usually know a set $U$ of possible values of the measurement error.

- This set may be a box, i.e., the set of all the tuples $(\Delta x_{j1}, \ldots, \Delta x_{jd})$ for which $\Delta^-_i \leq \Delta x_{ji} \leq \Delta^+_i$ for all $i$.
- This may be a subset of this box – e.g., an ellipsoid; see, e.g., [2–6, 11, 15, 16, 18, 20].

In all these cases, the set $U$ is a convex set containing 0.

An important aspect is that often, we do not know the set $U$ – i.e., we are not 100% sure about the measurement accuracy.

Let us summarize this information.

3.3 What we have and what we want

We have $n$ points $x_j = (x_{j1}, \ldots, x_{jd})$ in $d$-dimensional space. We know that there is a convex set $U$ containing 0 that describes measurement uncertainty.

We know the lower bound $k$ on the number of correct measurements. Usually, we know the proportion $\varepsilon$ of correct measurements, in this case $k = \varepsilon \cdot n$.

In precise terms, this means that there exists a point $x$ such that for at least $k$ of the original $n$ points, we have $x_j - x \in U$. Our objective is to generate a finite set $S$ – with as few points as possible – so that for one of the elements $s$ of this set, we have $s - x \in U$.

Let us describe this summary in precise terms.

3.4 Definition

**Definition 1.** Let $X = \{x_1, \ldots, x_n\}$ be a set of points in a $d$-dimensional space, and let $k < n$ be an integer. By the set $X$’s $(n, k)$-compression, we mean a finite set $S$ with the following property: For every pair $(U, x)$, for which:

- $U \in \mathbb{R}^d$ is a convex set containing 0,
- $x \in \mathbb{R}^d$, and
- $x_j - x \in U$ for at least $k$ different indices $j$,

there exists an element $s \in S$ for which $s - x \in U$. 


4 Main Result, Its Proof and Discussion

4.1 Result

Proposition 1. For every $d$, and for every $\delta > 0$, there exists a constant $c_{d,\delta}$ such that for all $\varepsilon$, if we take $k = \varepsilon \cdot n$, then for every set of $n$ points, there exists an $(n, k)$-compression with no more that $c_{d,\delta} \cdot \varepsilon^{-(d-0.5+\delta)}$ elements.

4.2 Discussion

Good news is that the above upper bound on the number of elements in a compression does not depend on $n$ at all.

- We can have thousands of measurement results,
- we can have billions of measurement results,

but no matter how many measurement results we have, we can always compress this information into a finite set whose size remains the same no matter what $n$ we choose.

4.3 Proof

This result follows from the known result of combinatorial convexity about so-called weak $\varepsilon$-nets; see, e.g., [1, 14]. A set $S$ is called a weak $\varepsilon$-net with respect to the set $X = \{x_1, \ldots, x_n\}$ if for every subset $Y$ of $X$ with at least $\varepsilon \cdot n$ elements, the convex hull of $Y$ contains at least element $s \in S$. It is known that for every dimension $d$ and for every number $\delta > 0$, there exists a constant $c_{d,\delta}$ such that for every set $X$, there exists a weak $\varepsilon$-net with $\leq c_{d,\delta} \cdot \varepsilon^{-(d-0.5+\delta)}$ elements.

To complete our proof, we need to show that each weak $\varepsilon$-net is an $(n, k)$-compression. Indeed, we know that for at least $k$ points, we have $x_j - x \in U$. Let us denote $k$ of these points by $x_{j_1}, \ldots, x_{j_k}$. In these terms, we have

$$x_{j_1} - x, \ldots, x_{j_k} - x \in U.$$

By the definition of the weak $\varepsilon$-net, one of the elements $s \in S$ belongs to the convex hull of the points $x_{j_1}, \ldots, x_{j_k}$, i.e., has the form $s = \alpha_1 \cdot x_{j_1} + \ldots + \alpha_k \cdot x_{j_k}$ for some values $\alpha_\ell \geq 0$ for which $\alpha_1 + \ldots + \alpha_k = 1$. Thus, we have

$$s - x = \alpha_1 \cdot (x_{j_1} - x) + \ldots + \alpha_k \cdot (x_{j_k} - x).$$

In other words, the difference $x - s$ is a convex combination of the differences $x_{j_1} - x$. The differences $x_{j_1} - x, \ldots, x_{j_k} - x$ all belong to the set $U$, and the set $U$ is convex, thus, there convex combination $s - x$ also belongs to $U$. The proposition is proven.
4.4 There exists an algorithm that computes the desired compression

Compressions are exactly weak $\varepsilon$-nets, and the problem of finding such a net can be described in terms of the first-order language of real numbers – with addition, multiplication, and inequalities, and finitely many quantifiers over real numbers. Thus, this problem is covered by an algorithm – this can be either the original Tarski-Seidenberg algorithm [17, 19] or one of its later improvements; see, e.g., [7, 13].

5 Remaining Open Problems

5.1 Can we get even smaller compressions?

Researchers in combinatorial convexity believe that we can have a weak $\varepsilon$-net of size $\leq c_d \cdot \varepsilon^{-1} \cdot (\log(1/\varepsilon))^C$ for some $C$ – and thus, a smaller-size compression – but this still needs to be proven.

5.2 How to efficiently compute a compression

While there exist algorithms for computing a small-size compression, these general algorithms require exponential time – even checking the condition that $s \in \text{Conv}(Y)$ for all subsets $Y$ of size $k$ requires checking exponential number of sets $Y$. Thus, for large $n$, these algorithms are not feasible. It is therefore desirable to come up with a feasible algorithm for computing the desired compression.

References