What If There Are Too Many Outliers?

Olga Kosheleva[1] and Vladik Kreinovich[2]

[1]Department of Teacher Education
[2]Department of Computer Science
University of Texas at El Paso
El Paso, Texas 79968, USA
olgak@utep.edu, vladik@utep.edu

1. What outliers we consider

- 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 an important case of what is known as "outliers".
- When we process measurement results, it is important to delete as many outliers as possible.

2. What outliers we consider

- Indeed, if we take outliers at face value:
  - we may get a biased 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 such outliers.

3. Problem

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

4. 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.

5. 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? How can we extract this information?

6. Why this extraction is not easy

- When we have a small number of outliers, we can:
  - delete them and
  - thus 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 exact 0s, 1/3 are 1s, and 1/3 are 2s.
- We know that no more than 2/3 of these results are outliers.
- 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.
- It could be that 2 is the actual value, and 0 and 1 are outliers.

7. Why this extraction is not easy (cont-d)

- 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.

8. Why this extraction is not easy (cont-d)

- Indeed, if 2 was the actual value, then we would have 5/6 outliers.
- This would contradict to our knowledge that the proportion of outliers does not exceed 2/3.

9. 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.
- Ideally, we should make this list as small as possible.

10. 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:
  - 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.

11. Often, a sensor measures several quantities (cont-d)

- In principle, we can simply consider such a complex measuring instrument as:
  - a collection of several instruments
  - that measure 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.
- So, 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.

12. Often, a sensor measures several quantities (cont-d)

- 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.

13. How can we represent uncertainty

- The measurement error $\Delta x \stackrel{\text{def}}{=} x_j - x$ is the difference between:
  - the measurement result x_j and
  - the (unknown) actual value x of the desired quantity.
- In the 1-D case, natural characteristics are the lower bound $\Delta^- < 0$ and the upper bound $\Delta^+ > 0$ on its value: $\Delta^- \le \Delta x_j \le \Delta^+$.
- In this case:
  - once we know the measurement result x_j,
  - we can conclude that the actual value x is located somewhere in the set $x_j - [\Delta^-, \Delta^+]$.
- Here, as usual, $x_j - U$ means the set of possible values $x_j - u$ when $u \in U$.
- In the general case, both x_j and x are d-dimensional, so the measurement error $\Delta x_j = x_j - x$ is also d-dimensional.

14. How can we represent uncertainty (cont-d)

- 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 $(\Delta x_{j1}, \ldots, \Delta x_d)$ for which $\Delta_i^- \le \Delta x_{ji} \le \Delta_i^+$ for all i.
- This may be a subset of this box – e.g., an ellipsoid.
- 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.

15. 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 ε 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 $x_j - x \in U$.
- We want 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.

16. Definition

- Let $X = \{x_1, \ldots, x_n\}$ be a set of points in a d-dimensional space, and let $k < n$ be an integer.
- We say that a pair (U, x) is X-*consistent* with X if:
  - $U \in \mathbb{R}^d$ is a convex set containing 0,
  - $x \in \mathbb{R}^d$, and
  - $x_j - x \in U$ for at least k different indices j,
- A set S is an (n, k)-*compression* of X if for every X-consistent pair (U, x), there exists en element $s \in S$ for which $s - x \in U$.

17. Result

- **Proposition.** *For every d, and for every $\delta > 0$, there exists a constant $c_{d,\delta}$ such that for all ε:*
  - *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.*

18. 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:
  - 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.

19. Proof

- This result follows from the known result of combinatorial convexity about so-called *weak ε-nets*.
- A set S is called a weak ε-net with respect to $X = \{x_1, \ldots, x_n\}$ 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 ε-net with $\leq c_{d,\delta} \cdot \varepsilon^{-(d-0.5+\delta)}$ elements.
- To complete our proof, we need to show that each weak ε-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 $x_{j_1}, \ldots, x_{j_k}$.

20. Proof (cont-d)

- In these terms, we have $x_{j_1} - x \in U$, $\ldots$, $x_{j_k} - x \in U$.
- By the definition of the weak ε-net, one of the elements $s \in S$ belongs to the convex hull of the points $x_{j_1}, \ldots, x_{j_k}$.
- So, $s = \alpha_1 \cdot x_{j_1} + \ldots + \alpha_k \cdot x_{j_k}$ for some values $\alpha_\ell \geq 0$ for which

$$\alpha_1 + \ldots + \alpha_k = 1.$$

- Thus, $s - x = \alpha_1 \cdot (x_{j_1} - x) + \ldots + \alpha_k \cdot (x_{j_k} - x)$.
- In other words, the difference $x - s$ is a convex combination of the differences $x_{j_\ell} - x$.
- 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.

21. There exists an algorithm that computes the desired compression

- Compressions are exactly weak ε-nets.
- So, 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 quantifies over real numbers.
- Thus, this problem is covered by an algorithm – this can be either the original Tarski-Seidenberg algorithm or one of its later improvements.

22. Can we get even smaller compressions?

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

23. 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.

24. References

- Bárány, I.: Combinatorial Convexity. American Mathematical Society, Providence, Rhode Island (2021)
- Rubin, N.: Stronger bounds for weak ε-nets in higher dimensions. Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing STOC 2021, Rome, Italy, June 21–25, 2021, pp. 989–1002 (2021)

25. Acknowledgments

- 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).
- It was also supported 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.