8-2016

One Needs to Be Careful When Dismissing Outliers: A Realistic Example

Carlos Fajardo
The University of Texas at El Paso, cmfajardo@miners.utep.edu

Olga Kosheleva
The University of Texas at El Paso, olgak@utep.edu

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

Comments:
Technical Report: UTEP-CS-16-54

One Needs to Be Careful When Dismissing Outliers: A Realistic Example

Carlos Fajardo[1], Olga Kosheleva[2], and Vladik Kreinovich[1]
[1]Department of Computer Science
[2]Department of Teacher Education
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
cmfajardo@miners.utep.edu,
olgak@utep.edu, vladik@utep.edu

Abstract

Traditional approach to eliminating outliers is that we compute the sample mean μ and the sample standard deviation σ, and then, for an appropriate value $k_0 = 2, 3, 6$, etc., we eliminate all data points outside the interval $[\mu - k_0 \cdot \sigma, \mu + k_0 \cdot \sigma]$ as outliers. Then, we repeat this procedure with the remaining data, eliminate new outliers, etc., until on some iteration, no new outliers are eliminated. In many applications, this procedure works well. However, in this paper, we provide a realistic example in which this procedure, instead of eliminating all outliers and leaving adequate data points intact, eliminates all the data points. This example shows that one needs to be careful when applying the standard outlier-eliminating procedure.

1 Formulation of the Problem

Need to eliminate outliers. In the traditional approach to data analysis, based on the sample, we estimate the means of the corresponding quantities, we estimate the variances, covariance, and correlations; see, e.g., [3]. This usually works well, but sometimes, we have *outliers*, i.e., values caused, e.g., by the malfunctioning of the measuring instrument.

Outliers ruin the estimations. For example, if we are interested in the average temperature, and in addition to 100 measurement results around 20° C, we have a (clearly erroneous) value 1000° C, then the sample average $\overline{x}$ becomes

$$\overline{x} \approx \frac{20 + \ldots + 20 \text{ (100 times)} + 1000}{101} \approx 30.$$

To get the means (and other statistical characteristics) that more adequately describe the situation, it is necessary to eliminate outliers before we start data processing.

How outliers are usually eliminated: main idea. A usual way to eliminate outliers is based on the fact that the most widely used normal distributions, the overwhelming majority of data points lie within the k_0-sigma interval

$$[\mu - k_0 \cdot \sigma, \mu + k_0 \cdot \sigma],$$

where μ is the mean, σ is the standard deviation, and the parameter k_0 determines our degree of confidence:

- for $k_0 = 2$, approximately 5% of values are outside the corresponding interval;
- for $k_0 = 3$, approximately 0.1% of the values are outside the interval;
- for $k_0 = 6$, approximately 10^{-8} of the values are outside the interval.

Thus, with a high probability, each value outside the corresponding k_0-sigma interval is an outlier.

This idea is widely used, e.g., in medicine. When medical doctors talk about the normal weight vs. overweight, what they mean is exactly this:

- someone with a weight in the k_0-sigma interval is considered normal, while
- someone with the weight outside this interval is considered to be overweight or underweight.

This is how normal blood pressure, normal height, and other characteristics are determined.

How outliers are usually eliminated: procedure. First, based on the sample $x_1, \ldots, x_n$, we estimate the sample mean $\mu = \dfrac{x_1 + \ldots + x_n}{n}$ and the sample standard deviation $\sigma = \sqrt{V}$, where $V = \dfrac{1}{n-1} \cdot \sum\limits_{i=1}^{n} (x_i - \mu)^2$. Then, for some pre-defined value k_0, we eliminate all the values x_i outside the interval $[\mu - k_0 \cdot \sigma, \mu + k_0 \cdot \sigma]$.

After that, the procedure is repeated again, to make sure that all the outliers are indeed eliminated. We stop when at some iterations, all the remaining values x_i are already within the corresponding k_0-sigma interval, i.e., when on this iteration, no new outliers are eliminated.

Example. In the above example, where the sample mean is $\mu \approx 30$, the sample variance is approximately equal to

$$V \approx \frac{1}{101} \cdot \left(100 \cdot 10^2 + 1000^2\right) \approx 10000,$$

thus $\sigma = \sqrt{V} \approx 100$, and so, e.g., the 3-sigma interval has the form

$$[30 - 3 \cdot 100, 30 + 3 \cdot 100] = [-270, 330].$$

Clearly, the value 1000 is outside this interval, so it is (correctly) classified as an outlier.

Why we sometimes need to repeat this procedure: a simple example. In the above example, one iteration of the above procedure was sufficient. So, why do we sometimes need to repeat this procedure?

To explain this need, let us assume that, in addition to 100 adequate measurement results, we have *two* outliers:

- the original outlier 1000 and
- a second outlier equal to 10^6.

In this case, after the first iteration of the above procedure, we get $\mu \approx 10^4$ and $\sigma \approx 10^5$. Thus, by considering everything outside a 3-sigma interval an outlier, we eliminate the largest outlier 10^6 – but the original outlier 1000 is still within the ranges.

To eliminate the original outlier, we need to repeat the same procedure again – in this case, as we have already explained, this outlier will be indeed eliminated.

What we do in this paper. In general, the above iterative procedure works well. However, we plan to provide a simple example where this procedure will not work at all – instead of eliminating all outliers, it will instead eliminate all the data.

The conclusion that we can make from this example is that one must be careful when applying a standard procedure for eliminating outliers.

2 Example

Example: main idea. As an example, let us consider a truncated power law distribution for which the probability density function (pdf) $\rho_0(x)$ is equal to 0 for $|x| > x_0$ and

$$\rho_0(x) = A_0 \cdot |x|^{-\alpha} \tag{1}$$

for $|x| \le x_0$ for some parameters A_0 and x_0.

Relations between parameters. The overall probability should be equal to 1, so we must have

$$\int_{-x_0}^{x_0} A_0 \cdot |x|^{-\alpha}\, dx = 1. \tag{2}$$

The pdf is symmetric with respect to the transformation $x \to -x$, so the integrals over $[0, x_0]$ and $[-x_0, 0]$ are equal to each other. Thus, the desired equality is equivalent to

$$\int_{0}^{x_0} A_0 \cdot |x|^{-\alpha} = 0.5. \tag{3}$$

This integral can be explicitly computed, so we get

$$A_0 \cdot \frac{x^{1-\alpha}}{1-\alpha}\Big|_0^{x_0} = 0.5. \tag{4}$$

To get a finite value for $x = 0$, we must have $1 - \alpha > 0$, i.e., $\alpha < 1$. In this case, the above requirement takes the form

$$A_0 \cdot \frac{x_0^{1-\alpha}}{1-\alpha} = 0.5, \tag{5}$$

hence

$$A_0 = \frac{1-\alpha}{2 \cdot |x_0|^{1-\alpha}}. \tag{6}$$

Let us apply the above procedure to this example: analysis of the problem. Let us analyze what happens if we apply the above outlier elimination procedure to this distribution.

In this procedure, we select some value k_0, and we dismiss all the values outside the interval $[\mu - k_0 \cdot \sigma, \mu + k_0 \cdot \sigma]$ as outliers. Since the selected distribution is symmetric with respect to 0, its mean is equal to $\mu = 0$. Due to the same symmetry, the integral over $[-x_0, x_0]$ computing the variance $V = \sigma^2$ is equal to twice the integral over the half-interval $[0, x_0]$:

$$V = \sigma^2 = \int_{-x_0}^{x_0} x^2 \cdot A_0 \cdot |x|^{-\alpha}\, dx = A_0 \cdot \int_{-x+0}^{x_0} |x|^{2-\alpha}\, dx =$$

$$2A_0 \cdot \int_0^{x_0} |x|^{2-\alpha}\, dx. \tag{7}$$

Integrating this expression and substituting the expression (6) for A_0, we conclude that

$$V = \sigma^2 = 2A_0 \cdot \frac{x_0^{3-\alpha}}{3-\alpha} = \frac{1-\alpha}{3-\alpha} \cdot x_0^2, \tag{8}$$

hence

$$\sigma = \sqrt{\frac{1-\alpha}{3-\alpha}} \cdot x_0. \tag{9}$$

Since here $\mu = 0$, the corresponding k_0-sigma interval $[\mu - k_0 \cdot \sigma, \mu + k_0 \cdot \sigma]$ has the form $[-k_0 \cdot \sigma, k_0 \cdot \sigma]$, i.e., the form $[-x_1, x_1]$, where $x_1 \stackrel{\text{def}}{=} k_0 \cdot \sigma$ has the form

$$x_1 = k_0 \cdot \sqrt{\frac{1-\alpha}{3-\alpha}} \cdot x_0. \tag{10}$$

If $x_1 \geq x_0$, then no value is dismissed. However, if $x_1 < x_0$, i.e., if

$$c \stackrel{\text{def}}{=} k_0 \cdot \sqrt{\frac{1-\alpha}{3-\alpha}} < 1, \tag{11}$$

then some values are dismissed.

After we dismiss all the values outside this interval, we get a new distribution, with the new probability density $\rho_1(x)$:

- which is equal to 0 when $|x| \geq x_1$ and
- which for $|x| \leq x_1$ has the form

$$\rho_1(x) = \rho_1(x \,|\, |x| \leq x_0) = \frac{\rho_1(x)}{\text{const}}, \tag{12}$$

where the constant in the denominator is the probability that for the original distribution, we have $|x| \leq x_1$.

Since the original pdf was proportional to $|x|^{-\alpha}$, we thus conclude that the new distribution has the following form:

- $\rho_1(x) = 0$ for $|x| > x_1$, and
- $\rho_1(x) = A_1 \cdot |x|^{-\alpha}$ for $|x| \leq x_1$, for an appropriate constant A_1.

In other words, the new distribution is similar to the original one, the only difference is that instead of the original cut-off x_0 we now have a new cut-off x_1 – which is determined by the formula $x_1 = c \cdot x_0$.

So what happens if we apply the above procedure again and again: a resulting description. We start with the power law distribution truncated by the cut-off values $x_0 > 0$.

After applying one iteration of the outlier elimination procedure, we get a similar power law distribution, but with a new cut-off value $x_1 = c \cdot x_0$, for some $c < 1$.

According to the general outlier elimination procedure, we have to continue these iterations until the process converges. Since the distribution $\rho_1(x)$ obtained after the first iteration has exactly the same form as the original distribution $\rho_0(x)$, we conclude that on the second iteration, we again get a similar distribution but with the new cut-off value $x_2 = c \cdot x_1 = c^2 \cdot x_0$.

On the third iteration, the cut-off value decreases to $x_3 = c \cdot x_2 = c^3 \cdot x_0$. On each iteration, we thus get a similar power-law distribution with cut-off values decreased by a factor of c: $x_{k+1} = c \cdot x_k$, hence $x_k = c^k \cdot x_0$.

As k increases, this sequence tends to 0. Thus, in this example, the above iterative process never stops: it eliminates all the non-zero values, and in the limit, we are left with only the value $x = 0$ – whose original probability is 0. Thus, if we apply the above procedure to this distribution, then, instead of eliminating all the outliers, we eliminate all the values altogether.

When can this happen? For this unexpected behavior to happen, we need to satisfy inequality (11), i.e., equivalently, the inequality

$$k_0 < \sqrt{\frac{3-\alpha}{1-\alpha}} \tag{13}$$

or, equivalently,

$$k_0^2 < \frac{3-\alpha}{1-\alpha} = 1 + \frac{2}{1-\alpha}. \tag{14}$$

This, in its turn, is equivalent to

$$\frac{2}{1-\alpha} > k_0^2 - 1, \tag{15}$$

i.e., equivalently,

$$\frac{1-\alpha}{2} < \frac{1}{k_0^2-1}$$

and

$$\alpha > 1 - \frac{2}{k_0^2-1}. \tag{16}$$

So:

- when $k_0 = 2$, this happens when $\frac{1}{3} < \alpha < 1$;
- when $k_0 = 3$, this happens when $\frac{3}{4} < \alpha < 1$;
- when $k_0 = 6$, this happens when $\frac{33}{35} < \alpha < 1$, with $\frac{33}{35} \approx 0.94$.

3 How Realistic Is This Example

Simplest possible distributions. To understand how realistic is the above example, let us consider the simplest possible distribution – the uniform distribution, in which each value has the same probability.

This distribution comes from the fact that often, the only information we have about a quantity y is that this quantity are the lower and upper bounds $\underline{y} \le y \le \overline{y}$. In principle, there are many probability distributions located on the interval $[\underline{y}, \overline{y}]$. Among all these distributions $\rho(y)$, it is reasonable to select the one for which the entropy $S = -\int \rho(y) \cdot \ln(\rho(y))\, dy$ is the largest possible; see, e.g., [2]. It is known that among all distributions located on an interval, the uniform distribution $\rho(y) = \text{const}$ has the largest possible entropy.

Selecting the uniform distribution makes perfect sense: if we have no reason to believe that some values y from the interval $[\underline{y}, \overline{y}]$ are more probable than others, then it makes sense to consider all these values equally probable – i.e., to consider them equally probable. This argument goes back to Laplace and is known as *Laplace's Indeterminacy Principle*.

The simplest case is when we have a random variable y uniformly distributed on the interval $[-1, 1]$.

From one random variable to another: reasonable transformations. In practice, instead of directly observing a variable y, we often observe an auxiliary variable x which is related to y by a known dependence $x = g(y)$.

What are the reasonable transformations $g(y)$? To decide which transformations are reasonable and which are not, let us take into account that the numerical value of a physical quantity depends on the choice of a measuring unit. If instead of the original measuring unit, we use a new one which

is λ times smaller, then all the numerical values increase by a factor of λ: $y \to y' = \lambda \cdot y$. For example, if we replace meters with centimeters, then 1.7 m becomes $100 \cdot 1.7 = 170$ cm.

In general, depending on the choice of the measuring unit, we have different numerical values describing the physical value. It is therefore reasonable to consider a transformation $x = g(y)$ to be physically reasonable if it does not depend on the choice of the measuring unit.

Of course, we cannot simply have $g(\lambda \cdot y) = g(y)$ for all x and λ, since this would simply imply that the function $g(y)$ is a constant. What we can require is that if we change a unit for measuring y, then after appropriately changing the unit for x, we get the exact same dependence. In other words, we require that for every λ, there exists a value $c(\lambda)$ for which $g(\lambda \cdot y) = c(\lambda) \cdot g(y)$ for all y.

It is known that under a natural condition of measurability, every function with this property has the form $g(y) = A \cdot y^\beta$ for some A and β; see, e.g., [1].

This holds for positive λ and y. For many physical quantities, negative values are also possible – e.g., for electric charge or for current or for a component v_x of the velocity. In all these cases, the choice of the sign is arbitrary: we could have easily selected as positive what we now call negative, and vice versa. In such situations, it is also reasonable to require that the reasonable transformation $x = g(y)$ be similarly invariant: e.g., if we simply change the sign for y, then after changing the sign for x, we get the exact same dependence. Under this requirement, for negative y, we get $g(y) = -g(|y|) = -A \cdot |y|^\beta$.

In general, we thus get $g(y) = \text{sign}(y) \cdot A \cdot |y|^\beta$.

What happens if we apply a reasonable transformation to the simplest uniform distribution. We want to know the probability density function $\rho_X(x)$ for which

$$\text{Prob}(x \le X \le x + dx) = \rho_X(x) \cdot dx.$$

For positive x, for which $X = g(Y) = A \cdot Y^\beta$ and thus, $Y = c \cdot X^B$, where we denoted $B \stackrel{\text{def}}{=} 1/\beta$ and $c \stackrel{\text{def}}{=} (1/a)^B$, the inequality $x \le X \le x + dx$ implies that

$$c \cdot x^B \le Y \le c \cdot (x + dx)^B.$$

Here,

$$(x + dx)^B = x^B + B \cdot x^{B-1} \cdot dx,$$

so

$$\rho_X(x) \cdot dx = \text{Prob}\left(c \cdot x^B \le Y \le c \cdot x^B + c \cdot B \cdot x^{B-1} \cdot dx\right).$$

The variable Y is uniformly distributed, so for this variable, the probability to be on a certain interval is proportional to the width of this interval. For our interval

$$[c \cdot x^B, c \cdot x^B + c \cdot B \cdot x^{B-1} \cdot dx],$$

the width is equal to $c \cdot B \cdot x^{B-1} \cdot dx$, thus $\rho_X(x) \cdot dx = \text{const} \cdot x^{B-1} \cdot dx$ and

$$\rho_X(x) \sim x^{B-1}.$$

For $x < 0$, we have a similar formula, so in general, we get $\rho_X(x) \sim |x|^{B-1}$. This is exactly our example of a distribution for which the standard way of eliminating outlier does not work – with $\alpha = 1 - B$. So, this example is indeed reasonable.

Comment. The unusual thing about our example distribution is that for this distribution, the probability density function $\rho(x)$ is unbounded. Can we have a similar example with a bounded continuous probability density function $\rho(x)$?

Not really. Indeed, the above feature means that as we use the above procedure to eliminate outliers, we get the distribution concentrated on an interval containing 0 that gets narrower and narrower – and its width tends to 0.

If the function $\rho(x)$ is continuous, this means that on the resulting small interval the probability density function get very close to constant – and thus, the distribution gets close to uniform. Each interval $[\underline{x}, \overline{x}]$ can be represented in the form $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$, where $\widetilde{x}$ is the interval's midpoint, and δ its half-width. For a uniform distribution on this interval, the mean is equal to $\widetilde{x}$, and the standard deviation is equal to $\sigma = \dfrac{1}{\sqrt{3}} \cdot \Delta$.

For $k_0 \geq 2$, we have

$$k_0 \cdot \sigma = k_0 \cdot \frac{1}{\sqrt{3}} \cdot \Delta \geq \frac{2}{\sqrt{3}} \cdot \Delta > \Delta.$$

Thus, all the values from the interval $[\underline{x}, \overline{x}] = [\widetilde{x} - \Delta, \widetilde{x} + \Delta]$ are inside the k_0-sigma interval $[\mu - k_0 \cdot \sigma, \mu + k_0 \cdot \sigma]$:

$$[\widetilde{x} - \Delta, \widetilde{x} + \Delta] \subseteq [\mu - k_0 \cdot \sigma, \mu + k_0 \cdot \sigma].$$

As a result, when the interval becomes sufficiently small, the above outlier-elimination procedure no longer eliminates any points – and therefore, for continuous bounded $\rho(x)$, we will never dismiss all the values (except for a single point).

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

[1] J. Aczél and J. Dhombres, *Functional Equations in Several Variables*, Cambridge University Press, 2008.

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

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