18.650: Fundamentals of Statistics November 8, 2023

Lecture 22 — Robust Statistics

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

1 Meaning of robust

Say we collect data

$$0.7, 0.1, -0.8, 275, -0.7, \ldots.$$

The number 275 is clearly an outlier. Formally, recall that we defined an outlier as any point X_i such that either $X_i > Q_3 + 1.5\text{IQR}$ or $X_i < Q_1 - 1.5\text{IQR}$, where IQR$= Q_3 - Q_1$. Note that this definition does *not* involve the mean or the standard deviation. This is because these two quantities are not robust, whereas the quantiles are robust.

Informally, we define

Definition 1.1: Robust statistics

Robust statistics are statistics which do not change in the presence of "a few" outliers.

Example.

> Suppose we have a dataset with 5 numbers between 0 and 1. The median is the third largest value. If we push any two of the 5 data points off to infinity, then three numbers will remain in the unit interval, and the median will necessarily be one of those three numbers. Therefore, the median stays bounded. On the other hand, if we move even a single data point off to infinity, then the *mean* of the five numbers will also go to infinity.

Motivated by this example, we quantify the robustness of an estimator using the following definition.

Definition 1.2: Breakdown point

Suppose we have a dataset $X_1, \ldots, X_n$ and an estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$, i.e. any function of the data. If the estimator $\hat{\theta}$ stays within a bounded interval after we replace up to $m-1$ of the observations by arbitrarily large or small outliers, and if this is no longer true if we replace m of the observations by outliers, then the breakdown point is $m/n \in [0, 1]$.

For example, suppose $\hat{\theta}(X_1, X_2, X_3, X_4, X_5)$ is the sample median. Let primes denote outliers. Then

$$|\hat{\theta}(X_1, X_2', X_3, X_4', X_5) - \hat{\theta}(X_1, X_2, X_3, X_4, X_5)| \le B < \infty$$

but

$$|\hat{\theta}(X_1, X_2', X_3', X_4', X_5) - \hat{\theta}(X_1, X_2, X_3, X_4, X_5)| \to \infty,$$

so the breakdown point is 3/5.

More generally, suppose we have a dataset of n numbers $X_1, \dots, X_n$ in the interval $[-B, B]$, where n is odd. Order the numbers as $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n+1)/2} \le \cdots \le X_{(n)}$. Suppose we replace $m = (n-1)/2$ of the numbers by outliers. The way to most significantly affect the data is to take the smallest $(n-1)/2$ numbers and send them to positive infinity, or to take the largest $(n-1)/2$ numbers and send them to negative infinity. In the first case, the median will change from $X_{(n+1)/2}$ to $X_{(n)}$. In the second case, the median will change from $X_{(n+1)/2}$ to $X_{(1)}$. Either way, the median stays bounded between $-B$ and B.

However, suppose instead we move the first $m = (n+1)/2$ smallest numbers off to infinity. Then we force the median to also move to infinity. Therefore ,the breakdown point of the median is $\frac{(n+1)/2}{n} = (n+1)/2n$, which converges to $1/2$ as $n \to \infty$.

Remark.

> Consider the estimator $\hat{\theta}(X_1, \dots, X_n) = 4$. You can replace any of the X_i without affecting the estimator, so the breakdown point is 1. However, for a *reasonable* estimator which actually tells us something about the data, a breakdown point of 1/2 is really the best we can do. This is because, when we go above 1/2, we start to have more outliers than non-outliers. The notion of "outlier" doesn't really make sense anymore if more than half the data is an outlier.

2 Median as MLE

We know the MLE has good statistical properties. But often, the MLE is given by the sample mean, which is not robust. Are there any distributions for which the *median* is the MLE — rather than the sample mean, as is most often the case?

Yes! This is true for the Laplace distribution, which has pdf

$$f_{\mu,b}(x) = \frac{1}{b} e^{-\frac{|x-\mu|}{b}}.$$

See Figure 1 for a plot of this density for different values of μ and b. For simplicity, let's fix $b = 1$. Our goal is to estimate the mean μ based on $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f_{\mu,1}$.

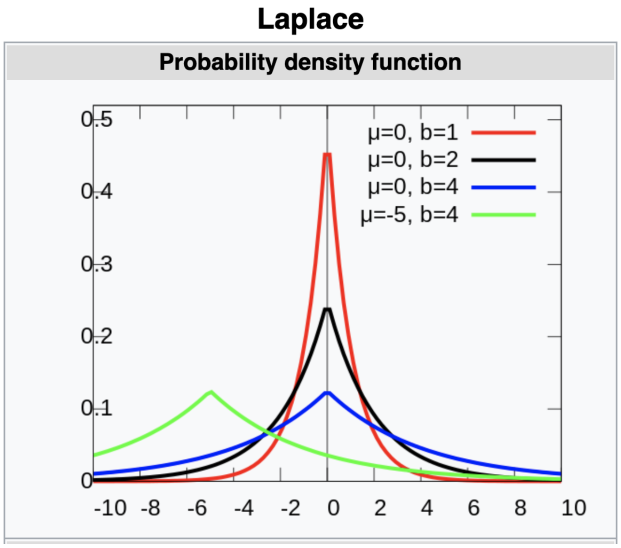

Figure 1: Density of the Laplace distribution.

The Laplace distribution has heavier tails than the Gaussian distribution. This means we are more likely to see outliers in the data, so the sample mean would not be such a good estimator of μ. And indeed, it turns out the MLE is the median of $X_1, \dots, X_n$. We have

$$\ell_n(\theta) = \sum_{i=1}^{n} \log(e^{-|X_i-\theta|}) = -\sum_{i=1}^{n} |X_i - \theta|.$$

So

$$\hat{\theta}^{\text{MLE}} = \text{argmin}_\theta \sum_{i=1}^{n} |X_i - \theta| = \text{median}(X_1, \dots, X_n).$$

3 Huber's contamination model and arbitrary contamination

Peter Huber is the father of robust statistics. To model the situation in which your data is "contaminated" by some outliers, he proposed the following mixture model.

We define the following three independent random variables:

$$Z \sim \text{Ber}(\epsilon), \quad X_{\text{true}} \sim \mathbb{P}_{\theta^*}, \quad X_{\text{out}} \sim Q,$$

where ϵ is some small number. Then, let Y be the random variable

$$Y = (1-Z)X_{\text{true}} + ZX_{\text{out}}.$$

In other words, with a small probability ϵ, we draw a sample from the outlier distribution Q. With probability $1-\epsilon$, we draw a sample from the true distribution. For example, if $\mathbb{P}_{\theta^*} = \mathcal{N}(\theta^*, 1)$ and Q has pdf q, then the pdf of Y is

$$f(y) = (1-\epsilon)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y-\theta^*)^2} + \epsilon q(y).$$

This gives us the following log likelihood:

$$\ell_n(\theta) = \sum_{i=1}^{n} \log\left((1-\epsilon)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(Y_i-\theta)^2} + \epsilon q(Y_i)\right).$$

However, the issue is that we need to know q to maximize ℓ_n. Rarely can we assume to know this outlier distribution. An alternative approach is to maximize over all possible q's in some family $\mathcal{Q}$:

$$\hat{\theta} = \text{argmax}_{\theta\in\Theta} \max_{q\in\mathcal{Q}} \ell_n(\theta). \tag{1}$$

This amounts to treating q as a nuisance parameter. The resulting estimator $\hat{\theta}$ is known as a quasi maximum likelihood estimator.

Remark.

> Another issue here is that we assumed that different samples drawn from the outlier distribution are independent of one another. This assumption may not always hold. This consideration motivates the next model.

3.1 Arbitrary contamination

Suppose all we know is that there is some set $\mathcal{C} = \{i \text{ s.t. } X_i \text{ is an outlier}\}$ of size $|C| = m$. We can then define a log likelihood which only incorporates the probabilities of those samples which are *not* outliers. For example, in the Gaussian case we would get

$$\ell_n(\theta, \mathcal{C}) = -\sum_{i\in\mathcal{C}^c}(X_i-\theta)^2.$$

We could then maximize

$$\hat{\theta} = \text{argmax}_{\theta} \max_{|\mathcal{C}|=m} \ell_n(\theta, \mathcal{C}).$$

One can check that the maximizer $\hat{\theta}$ removes the smallest $m/2$ X_i's as well as the highest $m/2$ X_i's, and takes the average of the remaining middle $n-m$ values. This is called a "*trimmed mean*". It works well in one dimension but is harder to even define in higher dimensions.