INTRODUCTION TO ROBUSTNESS: BREAKDOWN POINTS

Let $X = (X_1, ..., X_n)$ and $Z = (Z_1, ..., Z_n)$ be samples of real numbers. For $j = 1, ..., n$ let $X \ =_j \ Z$ mean that $X_i = Z_i$ except for at most j values of i. More specifically, for $y = (y_1, ..., y_j)$ let $X \ =_{j,y} Z$ mean that for some integers i_r with $1 \le i_1 < i_2 < ... < i_j \le n$, $Z_{i_r} = y_r$ for $r = 1, ..., j$ and $Z_i = X_i$ if $i \ne i_r$ for $r = 1, ..., j$. The idea is that X_i are i.i.d. from a nice distribution like a normal and y_r are errors or "bad" data. So the sample Z contains $n - j$ good data points and j errors. A robust statistical procedure will be one that doesn't behave too badly if j is not too large compared to n.

"Breakdown point" is one of the main ideas in robustness. Let $T = T(Z_1, ..., Z_n)$ be a statistic taking values in a parameter space Θ, a locally compact metric space. The main examples of parameter spaces to be considered here for real data are:

(a) The location parameter space of all μ such that $-\infty < \mu < \infty$ (the real line). Examples of statistics taking values in this space are the sample mean $\overline{Z}$ and the sample median.

(b) The scale parameter space containing 0, of all σ such that $0 \le \sigma < \infty$. Examples of statistics with values in $[0, \infty)$ are (i) the sample standard deviation and (ii) the median of all $|X_i - m|$ where m is the sample median. A variant of the scale parameter space is the open half-line $0 < \sigma < \infty$. Both examples (i) and (ii) can take the value 0 for some samples, so on such samples, these statistics are undefined if the scale parameter space is $(0, \infty)$.

(c) Often parameter spaces are considered, when location and scale are estimated simultaneously, of pairs (μ, σ) where $-\infty < \mu < \infty$ and $0 \le \sigma < \infty$ or alternately where $0 < \sigma < \infty$.

The closure of a set $A \subset \Theta$ will be denoted $\overline{A}$. If Θ is a Euclidean space or a closed subset of one, such as the closed half-line $0 \le \sigma < \infty$, then a set $A \subset \Theta$ has compact closure if and only if $\sup\{|x| : \ x \in A\} < \infty$. In the open half-line $0 < \sigma < \infty$, a subset A is compact if and only if it is bounded away from both 0 and $+\infty$, in other words for some $\delta > 0$ and $M < \infty$, $\delta \le \sigma \le M$ for all $\sigma \in A$.

The *breakdown point* of T at X is defined as

$$\varepsilon^*(T, X) \ = \ \varepsilon^*(T; X_1, ..., X_n) = \frac{1}{n} \max\{j : \ \overline{\{T(Z) : \ Z \ =_j \ X\}} \ \text{ is compact}\}.$$

In other words $\varepsilon^*(T, X) = j/n$ for the largest j for which there is some compact set $K \subset \Theta$ such that $T(Z) \in K$ whenever $Z \ =_j \ X$. If $\varepsilon^*(T, X)$ doesn't depend on X, which is often the case, then let $\varepsilon^*(T) \ := \ \varepsilon^*(T, X)$ for all X.

Some authors define the breakdown point instead in terms of the smallest number of replaced observations that can cause $T(Z)$ not to remain in any compact set. Such a definition adds $1/n$ to $\varepsilon^*(T, X)$ and makes no difference asymptotically as $n \to \infty$.

If a fraction of the data less than or equal to the breakdown point is bad (subject to arbitrarily large errors), the statistic doesn't change too much (it remains in a compact set), otherwise it can escape from all compact sets (in a Euclidean space, or by definition in other locally compact spaces, it can go to infinity). There are a number of definitions of breakdown point. The one just given is called the "finite sample" breakdown point (Hampel et al., 1986, p. 98, for a real-valued statistic).

Since j in the definition is an integer, the possible values of the breakdown point for samples of size n are $0, 1/n, 2/n, ..., 1$. A statistic with a breakdown point of 0 is (by definition) not robust. Larger values of the breakdown point indicate more robustness, up to breakdown point $= 1/2$ which is the maximum attainable in some problems.

Examples. (i) For the sample mean $T = \bar{Z} = (Z_1 + ... + Z_n)/n$, the breakdown point is 0 for any Z_j since for $j = 1$, if we let $y_1 \to \infty$ then $\bar{Z} \to \infty$ (for n fixed).
(ii) Let $T = Z_{(1)}$, the smallest number in the sample. Then the breakdown point of T is again 0 for any Z_i since for $j = 1$, as $y_1 \to -\infty$ we have $Z_{(1)} \to -\infty$. Likewise the maximum $Z_{(n)}$ of the sample has breakdown point 0.

So the statistics $\bar{Z}$, $Z_{(1)}$, $Z_{(n)}$ are not robust. Other order statistics have some robustness (for fixed finite n):

Theorem 1. For sample size n, and each $j = 1, ..., n$, the order statistic $T = Z_{(j)}$ has breakdown point $\varepsilon^*(T) = \frac{1}{n}\min(j-1, n-j)$.

Proof. At any sample $X = (X_1, ..., X_n)$, we have $\inf\{T(Z) : Z =_j X\} = -\infty$ (let $y_1, ..., y_j$ all go to $-\infty$). Likewise $\sup\{T(Z) : Z =_{n-j+1} X\} = +\infty$ (let $y_1, ..., y_{n-j+1} \to +\infty$). It follows that $\varepsilon^*(T, X) \leq \frac{1}{n}\min(j-1, n-j)$.

If $Z =_{j-1} X$ then the smallest possible value of $Z_{(j)}$ occurs when $y_i < X_k$ for all i and k and for at least one r such that $X_r = X_{(1)}$, X_r is not replaced, so $Z_{(j)} \geq X_{(1)}$. Similarly, if $Z =_{n-j} X$ the largest possible value of $Z_{(j)}$ satisfies $Z_{(j)} \leq X_{(n)}$. So if $k = \min(j-1, n-j)$ and $Z =_k X$, then $X_{(1)} \leq Z_{(j)} \leq X_{(n)}$ so $Z_{(j)}$ is bounded and $\varepsilon^*(T, X) = \frac{1}{n}\min(j-1, n-j)$ as claimed. Since this is true for an arbitrary X, the theorem is proved. □

If $j = 1$ or n, the breakdown point of $X_{(j)}$ is 0 as noted in the Examples above. If n is odd, so $n = 2k+1$ for an integer k, then the sample median $X_{(k+1)}$ has breakdown point $\frac{1}{2} - \frac{1}{2n} = \frac{k}{n}$. If $n = 2k$ for an integer k, then the two endpoints of the interval of medians, $Z_{(k)}$ and $Z_{(k+1)}$, each have breakdown point $\frac{1}{2} - \frac{1}{n}$. So any median has breakdown point at least $\frac{1}{2} - \frac{1}{n} \to \frac{1}{2}$ as $n \to \infty$. From Theorem 1, no other order statistic has any larger breakdown point than the median, so $\varepsilon^*(X_{(j)}) < 1/2$ for all j. This is typical behavior for interesting estimators. But, larger breakdown points are possible. If T has bounded values, then it trivially has breakdown point 1 by our definition. Or, let $T = \min_j |Z_j|$. Then one can check that T has breakdown point $1 - \frac{1}{n}$.

For real-valued observations $Z_1, \ldots, Z_n$, a real-valued statistic $T = T(Z_1, ..., Z_n)$ will be called *equivariant for location* if for all real θ, and letting $Z = (Z_1, \ldots, Z_n)$ and $Z + \theta = (Z_1 + \theta, ..., Z_n + \theta)$,

$$T(Z + \theta) = T(Z) + \theta$$

for all n-vectors Z of real numbers and all real θ.

For example, the order statistics $Z_{(j)}$ and the sample mean $\bar{Z}$ are clearly equivariant for location.

Theorem 2. For any real-valued statistic T equivariant for location, the breakdown point is $< 1/2$ at any $X = (X_1, ..., X_n)$.

Proof. Let the breakdown point of T at X be j/n. Then there is an $M < \infty$ such that

$$|T(Z)| \leq M \text{ whenever } Z \ =_j \ X. \tag{3}$$

Let $\theta = 3M$. Now $Z = Y + \theta$ for some Y with $Y \ =_j \ X$ if and only if $Z \ =_j \ X + \theta$. Then $T(Z) = T(Y) + \theta$. So

$$|T(Z) - \theta| \leq M \quad \text{whenever} \quad Z \ =_j \ X + \theta, \quad \text{and then} \quad 2M \leq T(Z) \leq 4M. \tag{4}$$

But if $j \geq n/2$ there is a Z with $Z \ =_j \ X$ and also $Z \ =_j X + \theta$. For such a Z, (3) and (4) give a contradiction, proving Theorem 2. □

REFERENCES

Frank R. Hampel, Peter J. Rousseeuw, Elvezio M. Ronchetti, and Werner A. Stahel (1986). *Robust Statistics: The Approach based on Influence Functions.* Wiley, New York.

Peter J. Huber (1981) *Robust Statistics.* Wiley, New York.