STAT 383C: Statistical Modeling I **Fall 2015**

Lecture 14 — October 13

Lecturer: Purnamrita Sarkar *Scribe: Some one*

14.1 Robust Statistics

We are now going to talk about statistics that are suitable when the data has outliers.

14.1.1 Some Definitions

Let us define the sample median by M_n and the sample mean by: $\bar{X}$. We know that: when $X_1 ... X_n$ are i.i.d. random variables drawn from a symmetric distribution with bounded variance, $\overline{X} \to \mu$. and furthermore via the Central Limit Theorem, $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \xrightarrow{d} N(0,1)$.

As it turns out, the median also converges to the population median, which in case of symmetric distributions is the mean. To be concrete, for the normal distribution, we have:

$$\frac{\sqrt{n}(M_n - \mu)}{\sigma} \xrightarrow{d} N(0, \pi/2)$$

Recall the ARE aka the asymptotic relative efficiency. It is the variance of the second argument divided by the variance of the first. i.e. $ARE(\overline{X}_n, M_n) = \pi/2 > 1$. Higher the ARE, we know that $\bar{X}_n$ is a better estimator than M_n for the normal, since it asymptotically has smaller variance, albeit they are both converging to the truth.

14.1.2 Some Examples of ARE

The Cauchy Lets take a Cauchy distribution with parameters μ and σ. Recall that a Cauchy has very heavy tails and in fact its mean and variance are undefined. So μ are σ here are the location and scale parameters respectively. A Cauchy distribution has very heavy tails. Suppose we have: $X_1, ... X_n \sim Cauchy(\mu, \sigma)$
In the case of a Cauchy distribution, the mean does not exist. What we have are:

$$\begin{aligned} \bar{X}_n &\sim Cauchy(\mu, \sigma^2) \\ M_n &\sim N(\mu, (\pi/4)^2 * \sigma^2/n) \\ ARE(\bar{X}_n, M_n) &= (\pi/4)^2 * 1/n \to 0 \end{aligned}$$

ν	$ARE(X_n, M_n)$
≤ 2	0
3	0.62
4	0.89
5	1.041

Table 14.1. ARE values as the degrees of freedom increases for a t distribution

So we see that for a heavy tailed symmetric distribution Median is a better estimator of the location parameter than the mean.

The t-distribution Lets now change our focus to the t distribution with ν degrees of freedom. For small ν the t distribution has heavy tails whereas the number of degrees of freedom increases, t converges to a normal distribution (see table 14.1).

14.1.3 Heuristic Approaches

We will start with some well known heuristics for estimating the location parameter when the data has outliers.

Definition 14.1. *α-**trimmed mean** The α trimmed mean orders the datapoints first and then trims a proportion α from both tails and calculates mean of the remaining datapoints.*

Definition 14.2. *α-**Winsorized mean** It removes α proportion of datapoints from the upper and lower tail and replaces them with the closest data points, i.e. the most extreme of the remaining data points.*

Lets do an example with the dataset $\{2, 4, 5, 10, 100\}$.

$$\begin{aligned} 20\% \text{ trimmed mean:} \quad & (4 + 5 + 10)/3 = 6.33 \\ 20\% \text{ Winsorized mean:} \quad & (4 + 4 + 5 + 10 + 10)/5 = 6.6 \end{aligned}$$

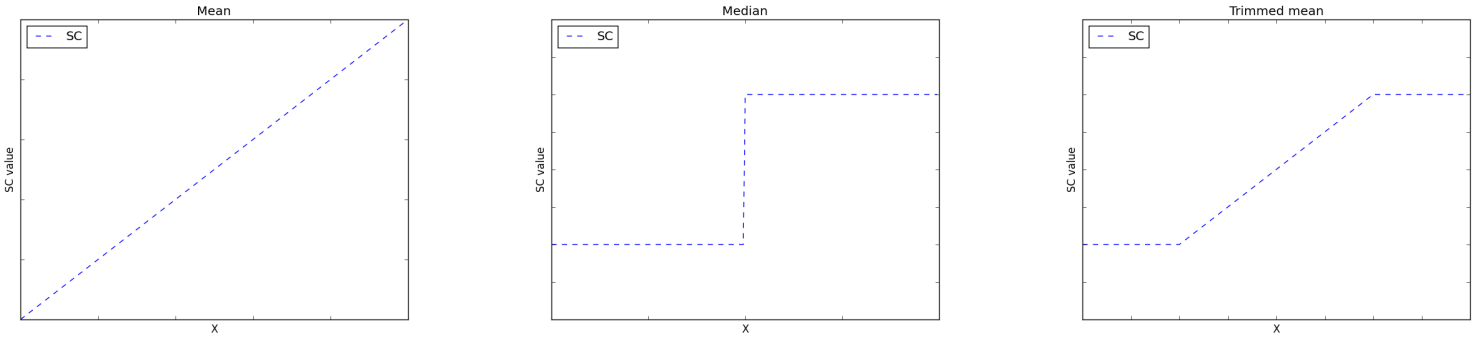
We will now formalize the notion of robustness.

14.1.4 Local and Global measures of robustness-

Definition 14.3. *The **Sensitivity curve** measures the effect of changing one datapoint on a given datapoint.*

$$SC_n(x, T) = \frac{T(X_1, ..., X_{n-1}, x) - T(X_1, ..., X_{n-1})}{(1/n)}$$

T is some sample statistic of interest, e.g., the median or mean. Intuitively, the sensitivity curve is analogous to a derivative of the statistic w.r.t the underlying distribution.



(A) (B) (C)

Table 14.2. In each of these figures, the midpoint of the x axis is the true location parameter. (A). Changing any element effects the mean. The larger the x-value of the removed element, the larger impact is made on the mean. (B). Now consider the case of the median. Adding or removing a single element shifts the median to an adjacent element. (C). In the case of α-trimmed mean, the sensitivity curve behaves the same as the sensitivity curve for the mean until we start removing elements that are within the αn elements at the tails that are discarded. At that point, the curve becomes flat.

Lets work out an example. Consider the sensitivity curve of the sample mean.

$$SC_n(x, \bar{X}_n) = \frac{((n-1)\bar{X}_{n-1} + x)/n - \bar{X}_{n-1}}{1/n} = x - \bar{X}_{n-1}$$

Definition 14.4. *The **Influence function** is:* $\lim_{n\to\infty} SC_n(x, T)$. *For example when* T_n *is the sample mean, we have* $IF(x, T, F) = x - T(F)$ *where* $T(F)$ *is the population mean (if indeed the mean exists). If the influence function of a statistic is bounded, then we call it robust.*

14.1.5 Breakdown Point

So far we have been looked at local notions of robustness. Now we will consider global notions of robustness. First we introduce the bias of a sample statistic as:

$$\text{bias}(m, T_n, Z) = sup_{Z'} |T_n(Z') - T_n(Z)|,$$

where Z' denotes a corrupted dataset obtained by replacing m datapoints in Z by arbitrary datapoints. Here, $sup_{Z'}$ can be thought of as a two player game, where you hand me m and I give you a corrupted dataset that can make your statistic as different as possible from the statistic computed from the uncorrupted dataset.

Definition 14.5. *The **Breakdown point** is defined as* $\epsilon^*(T_n, Z) = min\{m/n | bias(m, T_n, Z) = \infty\}$. *Intuitively, the breakdown point is how many points need to be replaced to make the bias go to* ∞.

For mean, it is asymptotically 0, for median it is 1/2, and for the α-trimmed mean it is asymptotically α.

14.1.6 M-estimators or maximum likelihood type estimators

Lets revisit ML estimation. We start with:

$$X_1 ... X_n \sim f(X, \theta)$$
$$L(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} f(X_i, \theta)$$
$$\hat{\theta} \leftarrow \arg\min \sum_{i=1}^{n} -\log(f(X_i, \theta))$$
$$\frac{\partial L}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log(f(X_i, \hat{\theta})) = 0$$

On the last line we are simply setting the score function to zero and solving for $\hat{\theta}$. As it turns out, M-estimators generalize MLE's. Instead of minimizing $-\log f(X_i; \theta)$ we will now minimize a loss function $\rho(X_i; \theta)$. Now define by $\psi(z) = d\rho(z)/dz$. This will replace the score function. So the necessary setup for M-estimation can be described by:

$$\hat{\theta} \leftarrow \arg\min \sum_{i=1}^{n} \rho(X_i, \theta)$$
$$\sum_{i=1}^{n} \psi(X_i, \hat{\theta}) = 0$$

$\rho(z)$ does not have to be tied to any particular log likelihood.

Examples When $\hat{\theta}$ is the sample mean, this is the squared loss $(X_i - \theta)^2$ whereas for the sample median it is the absolute loss $|X_i - \theta|$.

$$\rho(X, \theta) = (X - \theta)^2 \qquad \psi(X, \theta) = 2(X - \theta) \rightarrow \sum_i (X_i - \hat{\theta}) = 0 \rightarrow \hat{\theta} = \bar{X}_n$$
$$\rho(X, \theta) = |X - \theta| \qquad \psi(X, \theta) = \text{sign}(X - \theta) \qquad \text{when } X \neq \theta$$
$$\sum_i \text{sign}(X_i - \hat{\theta}) = 0 \rightarrow \hat{\theta} = \text{Median of } X_1, \ldots, X_n$$

How does all this tie in with what we have learned so far? It turns out that the influence function of an M estimator is proportional to its ψ function at point x. And so, as long as the $\psi(x)$ function is bounded for all x, we have a robust estimator.

14.1.7 Huber Loss

The famous Huber loss basically combines the squared loss and the absolute loss in an adaptive way.

$$\rho(X,\theta) = \begin{cases} 1/2(X-\theta)^2 & \text{If } |X-\theta| \leq c \\ c|X-\theta| - \frac{c^2}{2} & \text{o.w} \end{cases}$$

The corresponding ψ function is defined as:

$$\psi_H(X,\theta) = \begin{cases} X-\theta & \text{If } |X-\theta| \leq c \\ c \times \text{sign}(X-\theta) & \text{o.w} \end{cases}$$

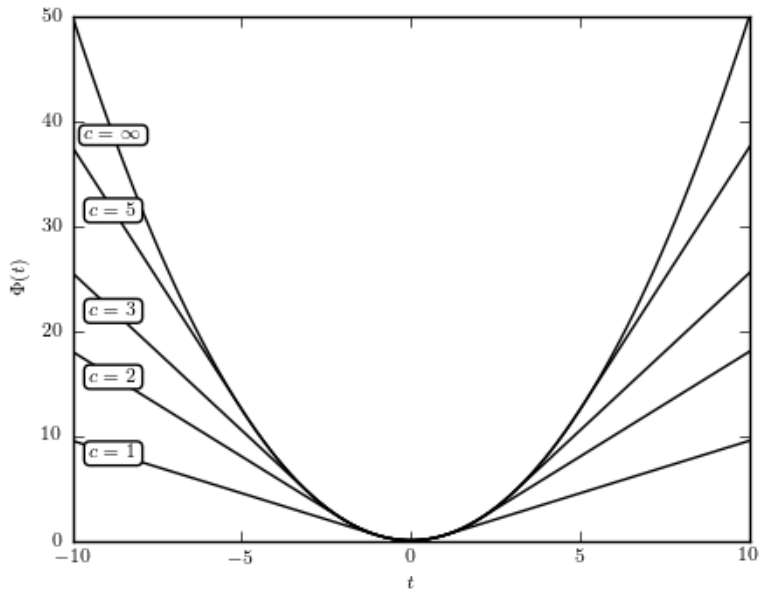


Figure 14.1. Shown above is Huber's loss function as the value of c ranges from 1 to ∞

Intuitively, when the data points are not too big, Huber's loss function penalizes like the squared loss, otherwise it penalizes like the absolute loss. c is a tuning parameter that is picked based on what the data is. Huber himself liked to set c to 1.345 which leads to 95% efficiency if $X_1 ... X_n \sim N(\mu, 1)$.

Up until now we have been pretending that we know the scale parameter or the variance. But what if that has to be estimated? Outliers will affect the variance far more than the mean, so we cannot use sample standard deviation. Instead we can use the MAD estimator: median absolute deviation. Actually is more like MADAM, median absolute deviation around the median.