Carnegie Mellon University

Robust Statistics

Vishwanath Saragadam Raja Venkata (vishwanath.srv@cmu.edu)

April 29, 2015

"A discordant small minority should never be able to override the evidence of the majority of the observations."

– Huber (2011)

Introduction

- Modeling of data most likely will deviate from the actual model.
- Experimental errors might crop up into data.
- Inference might be grossly wrong in that case.
- Can we come up with good statistics to capture these uncertainties in model?
- Is there a way to reduce the effect of outliers.

Why robust statistics

- Find an inference method that describes majority of the data.
- Identify outliers, i.e, data which does not fit the model.
- Talk about the influence of individual data points.
- Talk about how wrong the data has to be, to give a bad estimate.*Ronchetti,hem,Tyler*

What to expect from a robust statistic

- **Efficiency**: Reasonably good efficiency at the assumed mode.
- **Stability**: A small deviation from the assumed model shouldn't return garbage statistics.
- **Breakdown**: Large deviations shouldn't create a catastrophe.

Quantifying robustness – Sensitivity curve

- Let $T_n(\{x_i\}_{i=1}^n)$ be a statistic. Then the sensitivity curve of x,

$$SC(x, T) = \lim_{n\to\infty} n[T_n(x_1, \dots, x_{n-1}, x) - T_{n-1}(x_1, \dots, x_{n-1})] \tag{1}$$

- Quantifies the effect of an individual data point.

Quantifying robustness – Influence function

- Let F be a distribution and $F_\epsilon = (1-\epsilon)F + \epsilon\delta_x$ be the contaminated distribution.
- Let $T(F)$ be a statistic. Then,

$$IF(x, T, F) = \lim_{\epsilon \to 0} \frac{T(F_\epsilon) - T(F)}{\epsilon} = \frac{\partial}{\partial \epsilon} T(F_\epsilon)\Big|_{\epsilon=0} \tag{2}$$

- For mean, $IF(x, T, F) = x - \bar{\mu}$

Quantifying robustness – Breakdown point

- Define bias function:

$$bias(m, T_n, X) = \sup_{X'} \| T_n(X') - T_n(X) \| \tag{3}$$

Where X' is X with m points replaced by corrupted points.

- Breakdown point,

$$\epsilon_n^*(T_n, X) = \min\{\frac{m}{n} : bias(m, T_n, X) = \infty\} \tag{4}$$

- For mean, breakdown point is 0, because one rogue sample is sufficient to take bias to ∞

Some Ad-hoc ideas

- Agree that data comes from a normal distribution. This means, probability of outliers is very low. Solution? Drop such data points! Called α-trimmed mean.
- Another approach is to replace α proportion of tail data with it's closest observation. Called α-windsorized mean.
- Break down point of both methods is $\epsilon^* = \alpha$

The M-estimator

- Given data $\{x_i\}_{i=1}^n$ and statistic T_n.
- Assume that we wish to minimize the following function:

$$\min_{T_n} \sum_{i=1}^{n} \rho(x_i; T_n) \tag{5}$$

- Called an M-estimator, from Maximum likelihood type estimator.
- If we wish to find location, then $\rho(x_i; T_n) = \rho(x_i - T_n)$

The M-estimator

- Differentiating eq. 5,

$$\sum_{i=1}^{n} \psi(x_i; T_n) = 0 \tag{6}$$

- Eg, if $\rho(x_i; T_n) = \frac{1}{2}(x_i - T_n)^2, \psi(x_i; T_n) = (x_i - T_n)$. Simple least squares solution. Give sample mean.
- $\rho(x_i; T_n) = |x_i - T_n|, \psi(x_i; T_n) = sign(x_i - T_n)$. Gives median.

The M-estimator

- Intuitively, we want to penalize a large number of points with small error, but relax on a few points with large error.
- Huber loss function does this job.

$$h_k(x) = \begin{cases} \frac{1}{2}x^2 & |x| < k \\ k(|x| - \frac{1}{2}k) & |x| > k \end{cases} \tag{7}$$

- Breakdown point is 0.5, meaning that, more than 50% of the data has to be corrupt to give a bad estimate.

Robust statistic as an outcome of a PDF

- All methods described previously are based on some kind of intuition to deal with error.
- Can a robust estimate be an outcome of a density function.
- Heavy tail distributions have higher probability for tail end samples.
- Immediate distribution in mind: Cauchy distribution. Not reliable if sampling is truly normal.
- Can we get a control over the heaviness of the tail? Yes, student-t distribution Divgi (1990).

Using Student-t distribution for robust estimation

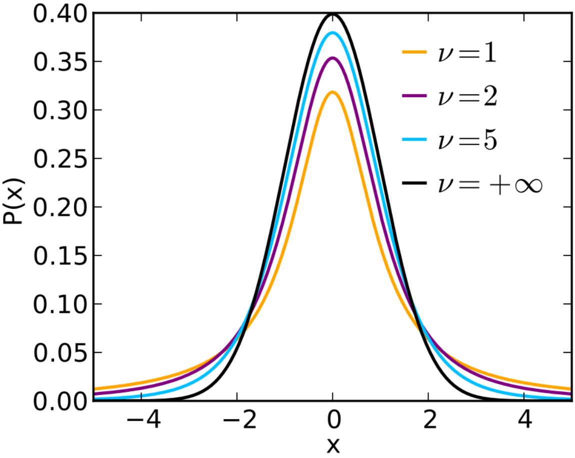

Figure : Image courtesy: Wikipedia.

Using Student-t distribution for robust estimation

- Let x come from a student-t distribution with center c, scale s and ν degrees of freedom.
- Let $u = \frac{x-c}{s}$. Then, the density of x,

$$f_X(x) = \frac{\left(1 + \frac{u}{\nu}\right)^{-\frac{\nu+1}{2}}}{s\sqrt{\nu}B(\frac{1}{2}, \frac{\nu}{2})} \tag{8}$$

- Log likelihood function for $\{x_i\}_{i=1}^n$,

$$L(c, s) = -\frac{\nu+1}{2}\sum_{i=1}^{n}\log(1 + \frac{u_i^2}{\nu}) - n\log(s\sqrt{\nu}B(\frac{1}{2}, \frac{\nu}{2})) \tag{9}$$

Using Student-t distribution for robust estimation

- Differentiating with respect to c,

$$\frac{\partial L}{\partial c} = 0 \implies \frac{\nu + 1}{s} \sum_{i=1}^{n} \frac{u_i}{\nu + u_i^2} = 0 \tag{10}$$

$$\equiv \sum_{i=1}^{n} \psi(x_i; \theta) = 0$$

- Differentiating with respect to s,

$$\frac{\partial L}{\partial s} = 0 \implies -\frac{n}{s} + \frac{\nu + 1}{s} \sum_{i=1}^{n} \frac{2u_i^2}{\nu + u_i^2} = 0 \tag{11}$$

Using Student-t distribution for robust estimation

- c, s can be estimated with gradient descent or alternating maximization algorithm.
- Tune ν for maximum log likelihood.
- Simple method, similar to M-estimator, and intuitive.
- Free of parameters.

Estimating mean using various methods

- Consider the data $\{x_i\}_{i=1}^{n}$, of which, m data points are corrupted.
- Add noise to $n - m$ data points, and perturb the m data points drastically.
- Try estimating mean of this data set using various methods.
- Vary m to see where each algorithm stops returning accurate mean.

Estimating mean using various methods

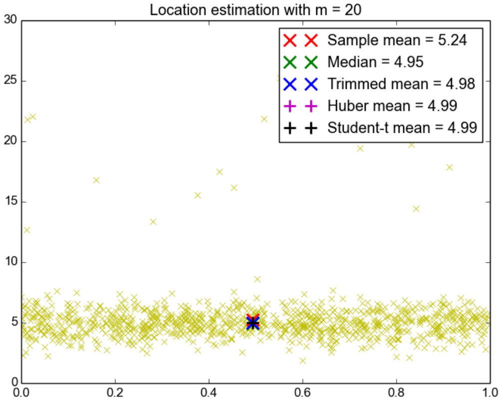

Figure : Location estimation with 20 outliers out of 1000

Estimating mean using various methods

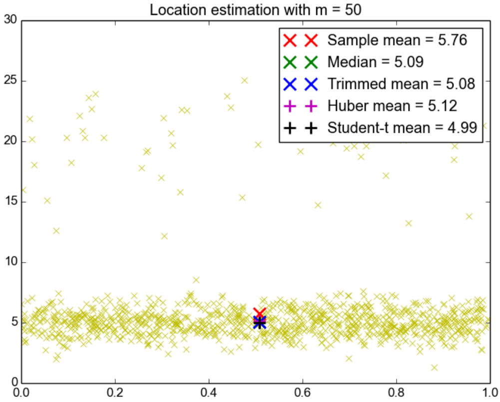

Figure : Location estimation with 50 outliers out of 1000

Estimating mean using various methods

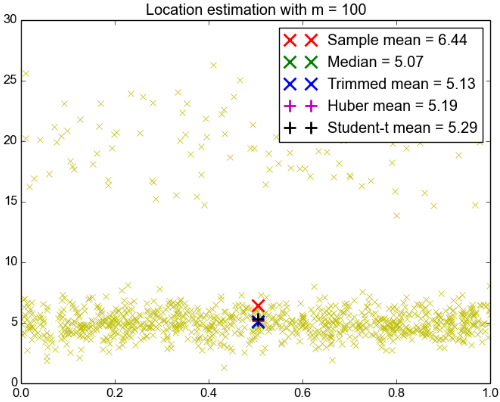

Figure : Location estimation with 100 outliers out of 1000

Estimating mean using various methods

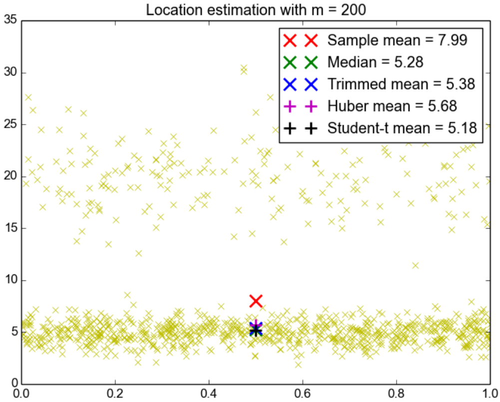

Figure : Location estimation with 200 outliers out of 1000

Estimating mean using various methods

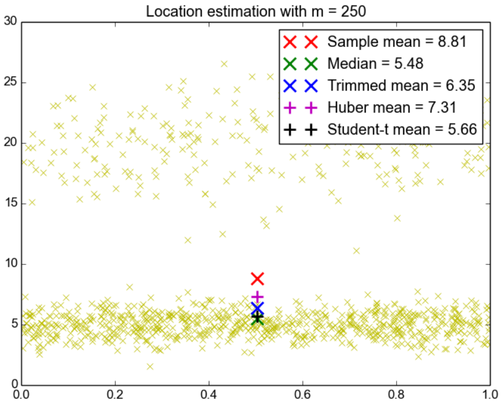

Figure : Location estimation with 250 outliers out of 1000

Concluding remarks

- Got a broad overview of robust statistics and it's necessity.
- Saw a couple of intuitive and well structured robust estimation techniques.
- No single best method for all problems. Need to go through some of the methods to figure out which one works.
- Many other robust estimation techniques like RANSAC, MINPRAN etc.

References

Robust statistics: a brief introduction and overview.

D R Divgi. Robust estimation using student's t distribution. 1990.

Peter J Huber. *Robust statistics*. Springer, 2011.

Elvezio Ronchetti. Introduction to robust statistics.

David E Tyler. A short course on robust statistics.