Investigating the Effectiveness of Robust Estimations in Multiple Dimensions

Afnan Kamran Khawaja, Banghua Zhu | TTE Presentation| UC Berkeley

ABSTRACT

- We are addressing the problem of natural outliers and data poisoning attacks in machine learning in which a small variation of data can contaminate the whole result. In this project, we will be investigating the effectiveness of different robust estimators by implementing and comparing the well-established high-dimensional robust estimators, including Tukey median.

BACKGROUND/GOALS

- Learning in the presence of outliers is an important goal in statistics and has been studied in the robust statistics community since the 1950's.
- $x_1, x_2, \ldots x_n \sim P$ where several samples are corrupted
- Estimate mean, given the corrupted data
- Validate/Estimate 1D mean and median
- Validate/Estimate 2D mean and Tukey median

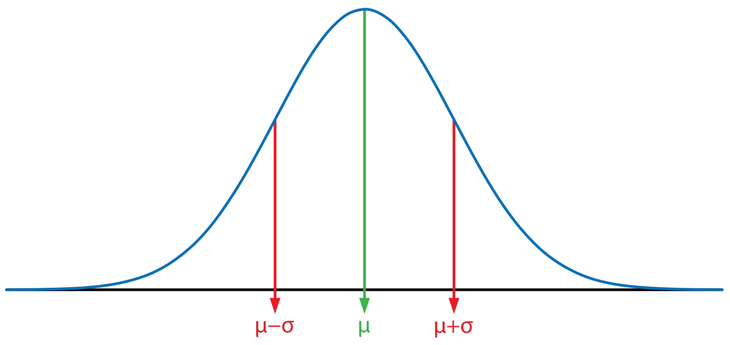
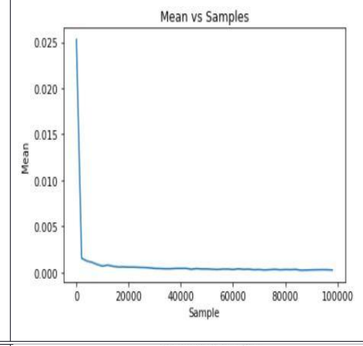
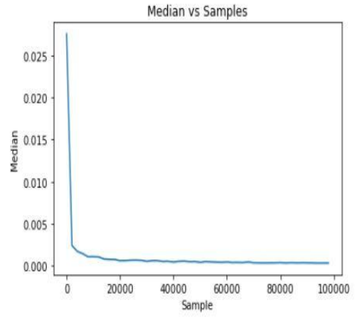
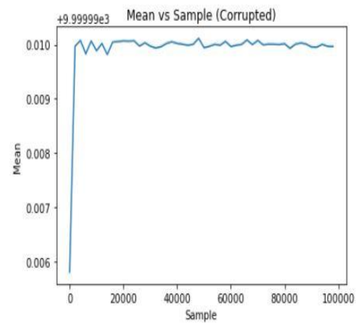
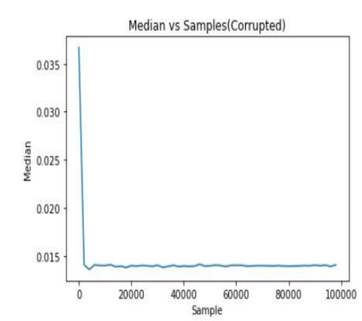
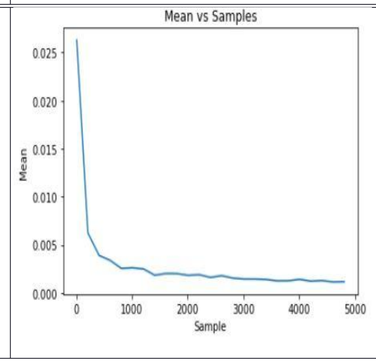
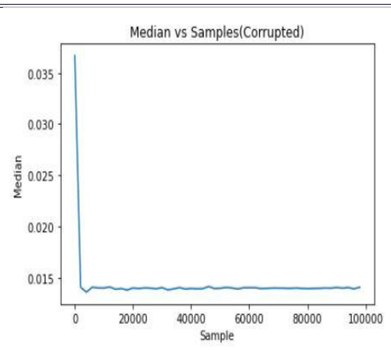
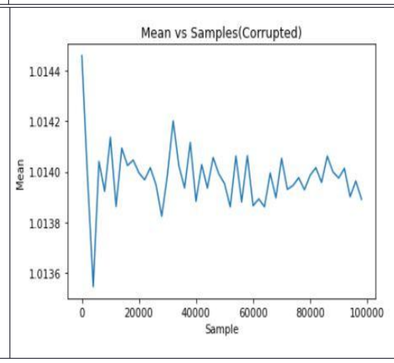
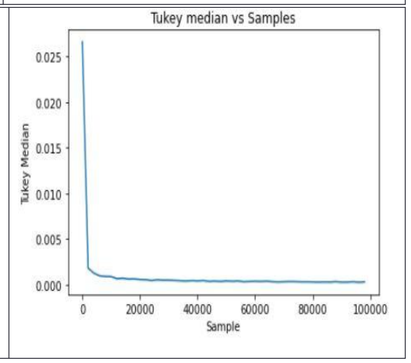


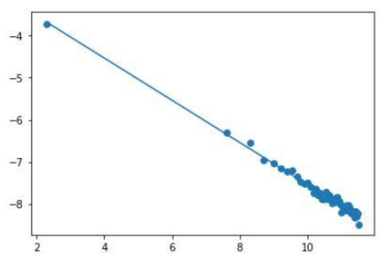
Fig. 1

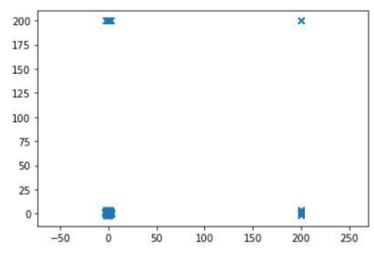
- A Gaussian distribution also known as the normal distribution is a very common probability distribution. Normal distributions are important in statistics and are used often when modeling real-valued random variable whose distributions are not known.

METHODS & RESULTS

	Non-Corrupted		Corrupted	
	Mean	Median	Mean	Median
1D				
2D				

CONCLUSION & FUTURE WORK





- We were able to get -2.54-0.5x where $\propto$=0.5 for the mean vs sample data points. For the median vs sample data set, we were able to get -2.52-0.5x where $\propto$=0.5. This showed us that the $1/\sqrt{x}$ is true for the data set.
- We concluded that the mean value is not usable when a data set is corrupted and that the median is more robust than mean.
- To compare it to the median, we would require the Tukey median which is in progress and if proved, would mean that median is more robust and multivariate data analysis as well as in singular dimensions.