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

METHODS & RESULTS

<table>
<thead>
<tr>
<th></th>
<th>Non-Corrupted</th>
<th>Corrupted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Me vs Samples</td>
<td>Mean vs Sample (Corrupted)</td>
</tr>
<tr>
<td>Median</td>
<td>Median vs Samples</td>
<td>Median vs Sample (Corrupted)</td>
</tr>
</tbody>
</table>

1D

2D

CONCLUSION & FUTURE WORK

• We were able to get \(-2.54-0.5x\) where \( \alpha = 0.5 \) for the mean vs sample data points. For the median vs sample data set, we were able to get \(-2.52-0.5x\) where \( \alpha = 0.5 \). This showed us that the \( 1/x \) 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.