

# Reviewing Robust Location and Scatter Estimators

---

Subhrajyoty Roy <sup>†</sup>

**Supervisors:** Prof. Ayanendranath Basu and Dr. Abhik Ghosh

February 17, 2022

<sup>†</sup> External Research Fellow  
Indian Statistical Institute, Kolkata

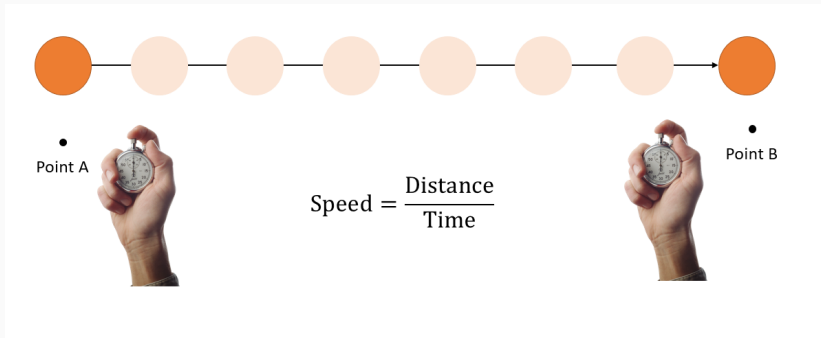
# OUTLINE

1. Introduction
2. Existing Robust Location and Covariance Estimators
3. Comparison of Existing Estimators
4. Application

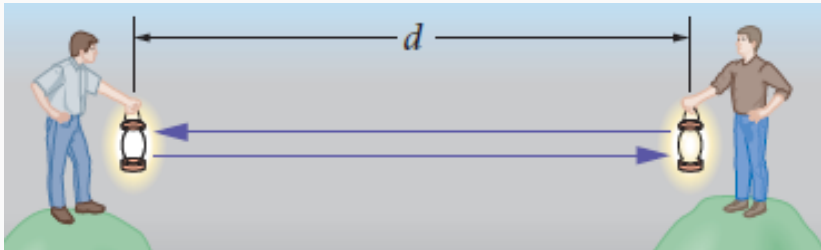
# Introduction

---

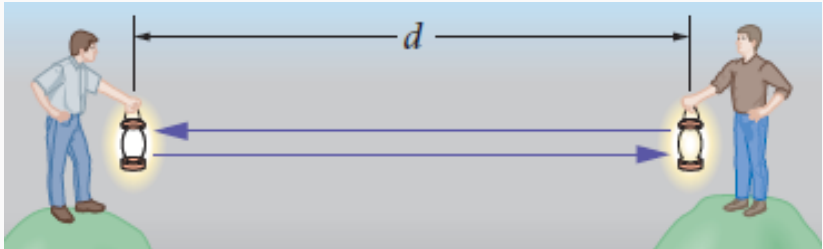
# A PROBLEM: MEASURING SPEED OF LIGHT



# GALILEO'S EXPERIMENT (1600s)



## GALILEO'S EXPERIMENT (1600s)

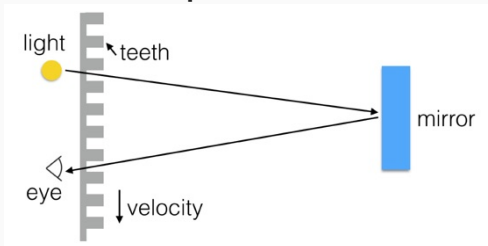


The actual explanation was not available until Einstein's general relativity on 1915.

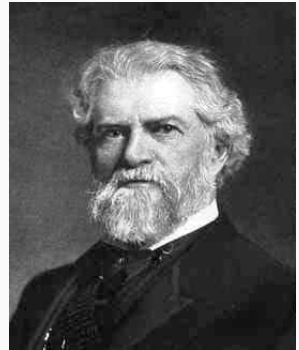
How do you ensure that the clocks are synchronized?

# NEWCOMB'S EXPERIMENT (1882)

## Newcomb's experiment:



- Eye on Fort Myers.
- Rotating mirror on Washington Tower.



Simon Newcomb  
(American Astronomer)

# NEWCOMB'S DATA

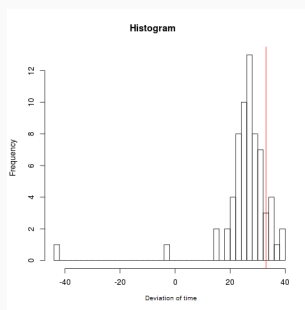
The data were recorded as deviation between amount of time it takes to travel 7442 meters for the light, and 24800 nanoseconds.

Since there were two obvious outliers, Newcomb discarded and used 10% trimmed mean to estimate speed of light.

**Newcomb's estimate:** 299,681,633 m/s.

**Currently accepted estimate:** 299,792,458 m/s.

**Error in Newcomb's estimate:** 0.037%.





# WHY?

1. Statistical models are built on assumptions.
2. Assumptions are often approximation of reality.
  - 2.1 Fuzzy knowledge.
  - 2.2 Outliers, part of data that differs significantly from majority of it.
3. Robust Statistical Inference builds methods that are resistant.
4. Achieves enough statistical guarantee even when assumptions fail to meet.

# How?

Huber defines three desirable features that every robust procedure should achieve.

1. **Stability** of the estimator under small deviations from assumed model.

# How?

Huber defines three desirable features that every robust procedure should achieve.

1. **Stability** of the estimator under small deviations from assumed model.
2. **Efficiency** under the assumed model.

# How?

Huber defines three desirable features that every robust procedure should achieve.

1. **Stability** of the estimator under small deviations from assumed model.
2. **Efficiency** under the assumed model.
3. **Breakdown**-resistance under large amount of contamination.

# GENERAL MODEL

Let,  $X_1, X_2, \dots, X_n$  be an i.i.d sample from  $F_\theta$ , for some unknown  $\theta \in \Theta$ .

We wish to make inference about  $\theta$ .

and let,  $T(X_1, \dots, X_n)$  be the proposed estimator of  $\theta$ .

We denote  $T : \mathcal{F} \rightarrow \mathbb{R}$  as the corresponding functional, i.e.,

$$T(F_n) = T(X_1, \dots, X_n), \quad T(F_\theta) = \theta$$

# GENERAL MODEL

Let,  $X_1, X_2, \dots, X_n$  be an i.i.d sample from  $F_\theta$ , for some unknown  $\theta \in \Theta$ .

We wish to make inference about  $\theta$ .

and let,  $T(X_1, \dots, X_n)$  be the proposed estimator of  $\theta$ .

We denote  $T : \mathcal{F} \rightarrow \mathbb{R}$  as the corresponding functional, i.e.,

$$T(F_n) = T(X_1, \dots, X_n), \quad T(F_\theta) = \theta$$

## Examples:

1. For sample mean,  $T_1(F) = \int x dF$ .
2. For sample median,  $T_2(F) = F^{-1}(1/2)$ .

# INFLUENCE FUNCTION

Let,  $F_\epsilon = (1 - \epsilon)F_\theta + \epsilon\delta_x$ .

Let,  $Y_1, \dots, Y_n \sim F_\epsilon$ .

$$\text{Empirical } IF(x; T, F_\theta) = \lim_{\epsilon \rightarrow 0+} \frac{T(Y_1, \dots, Y_n) - T(X_1, \dots, X_n)}{\epsilon}$$

In terms of functional,

$$IF(x; T, F) = \lim_{\epsilon \rightarrow 0+} \frac{T((1 - \epsilon)F + \epsilon\delta_x) - T(F)}{\epsilon}$$

# ASYMPTOTIC VARIANCE

Using Von Mises expansion, one can write

$$T(F_n) = T(F) + \int IF(x; T, F) dF_n(x) + \text{remainder}$$

$$\Rightarrow \sqrt{n}(T(F_n) - T(F)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n IF(X_i, T, F) + \text{remainder}$$

Under standard regularity assumptions,

$$\Rightarrow \sqrt{n}(T(F_n) - T(F)) \xrightarrow{d} \mathcal{N}\left(0, \int IF(x; T, F)^2 dF(x)\right)$$

Therefore, the asymptotic variance of centered and normalized estimator  $T(F_n)$  is

$$V(T, F) = \int IF(x; T, F)^2 dF(x)$$



## BREAKDOWN POINT

**Example:** Given a sample  $X_1, \dots, X_n$  and the estimator sample mean  $\bar{X}$ , one can modify  $X_1$  to make  $\bar{X}$  as large (or as small) as required. In other words, one can break its reliability by contaminating only one sample.

## BREAKDOWN POINT

**Example:** Given a sample  $X_1, \dots, X_n$  and the estimator sample mean  $\bar{X}$ , one can modify  $X_1$  to make  $\bar{X}$  as large (or as small) as required. In other words, one can break its reliability by contaminating only one sample.

$$\epsilon_n^*(T; x_1, x_2, \dots, x_n) = \frac{1}{n} \max \left\{ m : \max_{i_1, \dots, i_m} \sup_{Y_1, \dots, Y_m} T(Z_1, Z_2, \dots, Z_n) < \infty \right\}$$

where  $Z_l = X_l$  if  $l \notin \{i_1, \dots, i_m\}$  and  $Z_l = Y_j$  if  $l = i_j$ .

**Breakdown point** is the limit of  $\epsilon_n^*$  as  $n \rightarrow \infty$ .

## **Existing Robust Location and Covariance Estimators**

---

# MODEL

Let,  $X_1, X_2, \dots, X_n$  be an i.i.d sample  $\sim F(\mu, \Sigma)$ , each  $X_i \in \mathbb{R}^p$ .

Also, assume  $\mathbb{E}(X_i) = \mu$  and  $\text{Var}(X_i) = \Sigma$ .

Both  $\mu$  and  $\Sigma$  are unknown.

We want to estimate these robustly, so that even if some of the  $X_i$ s come from  $G$ , such that  $G \approx F$ , our estimate won't change rapidly.

# TRIMMED AND WINSORIZED ESTIMATOR

Let,  $z_1, z_2, \dots, z_n$  be univariate samples.

Since, usual mean is nonrobust, as it has unbounded influence function.

We wish to have a bounded influence for each sample  $z_i$ .

Let,  $z_{(1)} < z_{(2)} < \dots < z_{(n)}$  be the order statistics.

1.  $\alpha$ -Trimmed Estimator:

$$\bar{z}_\alpha = \frac{1}{(n - 2[n\alpha])} \sum_{i=(1+[n\alpha])}^{n-[n\alpha]} z_{(i)}$$

2.  $\alpha$ -Winsorized Estimator:

$$\bar{z}_\alpha^* = \frac{1}{n} \left( [n\alpha]z_{(1+[n\alpha])} + \sum_{i=(1+[n\alpha])}^{n-[n\alpha]} z_{(i)} + [n\alpha]z_{(n-[n\alpha])} \right)$$

# MULTIVARIATE TRIMMED AND WINSORIZED ESTIMATOR

Concept of order statistic is complicated! Requires data depth! Bickel (1965) introduced two multivariate estimators in similar direction.

1.  $\lambda$ -Trimmed Estimator:

$$\bar{X}_\lambda(\hat{\theta}) : \hat{\theta} = \frac{1}{\sum_{i=1}^n \mathbf{1}(\|X_i - \hat{\theta}\| < \lambda)} \sum_{i=1}^n X_i \mathbf{1}(\|X_i - \hat{\theta}\| < \lambda)$$

2.  $\lambda$ -Winsorized Estimator:

$$\bar{X}_\lambda^*(\hat{\theta}) : \hat{\theta} = \frac{1}{n} \left( \sum_{i=1}^n \left( \hat{\theta} + \lambda \frac{(X_i - \hat{\theta})}{\|X_i - \hat{\theta}\|} \right) \mathbf{1}(\|X_i - \hat{\theta}\| \geq \lambda) + \sum_{i=1}^n X_i \mathbf{1}(\|X_i - \hat{\theta}\| < \lambda) \right)$$

Scatter estimates follow from sample covariance matrix of trimmed / winsorized samples.

# M-ESTIMATOR

Consider MLE, we wish to maximize the log likelihood

$$\max_{\mu, \Sigma} n^{-1} \sum_{i=1}^n \ell(\mu, \Sigma; X_i).$$

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial \ell}{\partial \mu}(X_i) = 0$$

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial \ell}{\partial \Sigma}(X_i) = 0$$

Here, each sample  $X_i$  has the same contribution, irrespective of its deviation from the model.

# M-ESTIMATOR

Score  $\ell'(X)$  is unbounded in  $X$ , hence any one sample  $X_i$  can make things problematic!

We need to introduce bounded functions!



# M-ESTIMATOR

Score  $\ell'(X)$  is unbounded in  $X$ , hence any one sample  $X_i$  can make things problematic!

We need to introduce bounded functions!

General estimating equation,

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial \rho}{\partial \mu}(X_i) = \frac{1}{n} \sum_{i=1}^n \frac{\partial \rho}{\partial \Sigma}(X_i) = 0$$

If  $\rho'$  are bounded, then effect of every sample  $X_i$  is bounded, hence no one point can influence the estimator to behave erratically.

# M-ESTIMATOR

Maronna (1976) proposed  $M$ -estimators of location and scatter from the same idea, as the solution of the estimating equations

$$\frac{1}{n} \sum_{i=1}^n \Psi_1 \left( (X_i - \hat{\mu})^\top \hat{\Sigma}^{-1} (X_i - \hat{\mu}) \right) (X_i - \hat{\mu}) = 0$$

$$\frac{1}{n} \sum_{i=1}^n \Psi_2 \left( (X_i - \hat{\mu})^\top \hat{\Sigma}^{-1} (X_i - \hat{\mu}) \right) (X_i - \hat{\mu})(X_i - \hat{\mu})^\top = \hat{\Sigma}$$

where  $\Psi_1, \Psi_2$  are two bounded functions with both decreasing in absolute value of its argument.

# DESIRABLE PROPERTIES OF ROBUST LOCATION ESTIMATE

Rousseeuw (1985) described two desirable properties for estimator of location for a multivariate sample.

Let,  $T(X_1, \dots, X_n)$  be the estimator of location of the samples  $X_1, \dots, X_n$ , each  $X_i \in \mathbb{R}^p$ .

1. Breakdown point  $\epsilon(T, X)$  should be large, "close" to  $1/2$ .

# DESIRABLE PROPERTIES OF ROBUST LOCATION ESTIMATE

Rousseeuw (1985) described two desirable properties for estimator of location for a multivariate sample.

Let,  $T(X_1, \dots, X_n)$  be the estimator of location of the samples  $X_1, \dots, X_n$ , each  $X_i \in \mathbb{R}^p$ .

1. Breakdown point  $\epsilon(T, X)$  should be large, "close" to  $1/2$ .
2. It should be affine equivariant, i.e., for any  $b \in \mathbb{R}^p$  and nonsingular matrix  $A$ ,

$$T(AX_1 + b, \dots, AX_n + b) = AT(X_1, \dots, X_n) + b$$

# THREE COMPETING ESTIMATORS?

## Sample mean $\bar{X}$

1. Breakdown at 0.
2. Affine equivariant.

# THREE COMPETING ESTIMATORS?

$L_1$  median

Sample mean  $\bar{X}$

$$\min_{a \in \mathbb{R}^p} \sum_{i=1}^n \|X_i - a\|_{L_1}$$

1. Breakdown at 0.
2. Affine equivariant.

1. Breakdown at  $1/2$ .
2. Not affine equivariant.

# THREE COMPETING ESTIMATORS?

## $L_1$ median

### Sample mean $\bar{X}$

1. Breakdown at 0.
2. Affine equivariant.

$$\min_{a \in \mathbb{R}^p} \sum_{i=1}^n \|X_i - a\|_{L_1}$$

1. Breakdown at  $1/2$ .
2. Not affine equivariant.

### M-estimator

1. Breakdown at  $1/(p+1)$ .
2. Affine equivariant.

# MVE ESTIMATOR

No! we are not trading affine equivariance and breakdown.

Define,

$T(X_1, \dots, X_n)$  = Center of the minimum volume ellipsoid  
containing at least  $h$  points of  $X_1, \dots, X_n$

Usually,  $h = \lfloor n/2 \rfloor + 1$ .



# MVE ESTIMATOR

No! we are not trading affine equivariance and breakdown.

Define,

$T(X_1, \dots, X_n)$  = Center of the minimum volume ellipsoid  
containing at least  $h$  points of  $X_1, \dots, X_n$

Usually,  $h = \lfloor n/2 \rfloor + 1$ .

1. Clearly, affine equivariant.
2. Breakdown at  $(\lfloor n/2 \rfloor - p + 1)/n \approx 1/2$  for large  $n$ .
3. Cannot be computed if  $p > \lfloor n/2 \rfloor + 1$ .
4. NP-hard to solve.

# MCD ESTIMATOR

Instead, we consider the confidence ellipsoid's volume, for Gaussian distribution. Define,

$T(X_1, \dots, X_n)$  = Mean of the  $h$  points among  $X_1, \dots, X_n$   
such that determinant of covariance matrix is minimal.

Usually,  $h = \lceil n/2 \rceil + 1$ .

# MCD ESTIMATOR

Instead, we consider the confidence ellipsoid's volume, for Gaussian distribution. Define,

$T(X_1, \dots, X_n)$  = Mean of the  $h$  points among  $X_1, \dots, X_n$   
such that determinant of covariance matrix is minimal.

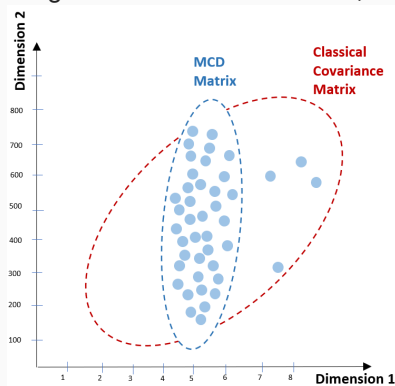
Usually,  $h = \lfloor n/2 \rfloor + 1$ .

1. Clearly, affine equivariant.
2. Breakdown at  $(\lfloor n/2 \rfloor - p + 1)/n \approx 1/2$  for large  $n$ .
3. Cannot be computed if  $p > \lfloor n/2 \rfloor + 1$ .
4. Easier to solve by taking convex hull.

# LOCATION AND SCATTER ESTIMATOR

Once we identify the best  $h$  points through MVE or MCD estimator,

1. Robust estimate of location is sample mean of those best  $h$  points.
2. Robust estimate of scatter is sample covariance of those best  $h$  points.



# S-ESTIMATOR

Introduced by Rousseeuw (1984), develops from a regression problem.

Assume that the data comes from a distribution  $F$ ,  $X_1, \dots, X_n \sim F$ .

$$\frac{1}{n} \sum_{i=1}^n E_F \left[ \left( \frac{X_i - \mu}{\sigma} \right)^2 \right] = 1$$

For a robust estimate, we might want

$$\frac{1}{n} \sum_{i=1}^n E_F \left[ \left| \frac{X_i - \mu}{\sigma} \right| \right] = K'$$

In general, we have scale estimator.

$$\frac{1}{n} \sum_{i=1}^n E_F \left[ \rho \left( (X_i - \mu)^\top \Sigma^{-1} (X_i - \mu) \right) \right] = K$$

# S-ESTIMATOR

The function  $\rho(\cdot)$  satisfy,

1.  $\rho(0) = 0$ .
2.  $\rho(\cdot)$  is symmetric about 0 and is increasing in magnitude.
3. It is bounded.

Davies (1987) improvised it to an optimization problem

$$\text{Minimize } \det(\Sigma) \text{ subject to } \sum_{i=1}^n \rho((X_i - \mu)^\top \Sigma^{-1} (X_i - \mu)) \geq K$$

it has much better properties than Rousseeuw's version of S-estimator.

# STAHEL-DOHONO ESTIMATOR

Usual nonrobust location and scatter estimators,

$$\hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^n X_i$$

and

$$\hat{\Sigma}_1 = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu}_1)(X_i - \hat{\mu}_1)^\top$$

The problem is each  $X_i$  has same influence, irrespective of its deviation from the model.

# STAHEL-DOHONO ESTIMATOR

Idea is to introduce weights!

$$\hat{\mu}_2 = \frac{\sum_{i=1}^n w(X_i) X_i}{\sum_{i=1}^n w(X_i)}$$

and

$$\hat{\Sigma}_2 = \frac{\sum_{i=1}^n w(X_i) (X_i - \hat{\mu}_2)(X_i - \hat{\mu}_2)^T}{\sum_{i=1}^n w(X_i)}$$

Here,  $w(X_i)$ s are such that for points with high deviation from the assumed model, they are small.



# STAHEL-DOHONO ESTIMATOR

## How to choose these weights?

Look for some one-dimensional direction in which  $X_i$  is most outside from the data cloud.

$$w(X_i) \propto \text{Maximum value of } \mathbf{u}^\top X_i \text{ subject to } \|\mathbf{u}\| = 1$$

Usually, we should center and scale this before projecting onto  $\mathbf{u}$ ,

$$w(X_i) = \sup_{\|\mathbf{u}\|=1} \frac{\mathbf{u}^\top X_i - \text{med}_{1 \leq j \leq n} \mathbf{u}^\top X_j}{\text{med}_k |\mathbf{u}^\top X_k - \text{med}_{1 \leq j \leq n} \mathbf{u}^\top X_j|}$$

# **Comparison of Existing Estimators**

---

# COMPARISON OF ESTIMATORS

Method	Affine Equivariance	Asymptotic Breakdown Point	Asymptotic Property	Computational Complexity	Assumption on dimensionality
W-estimator	Yes	$\alpha$	$\sqrt{n}$ -consistent Asymptotic Normal Less efficient for high $\alpha$	$O(np^3)$ / iteration	None
MVE	Yes	1/2	$n^{1/3}$ -consistent Not asymptotic normal	NP-hard Not known	$p \leq [n/2] + 1$
MCD	Yes	1/2	$\sqrt{n}$ -consistent Asymptotic Normal Not very efficient	$O(n^{\min(p^2, h)} \log(n))$	$p \leq [n/2] + 1$
M-estimator	Depends on $\Psi(\cdot)$	$\leq \frac{1}{(p+1)}$ , if AE Depends on $\Psi(\cdot)$ Could be 1/2 for some $\Psi$	$\sqrt{n}$ -consistent Asymptotic Normal Very efficient	$O(np^3 f(n, p))$ / iteration	Usually $p < n$
S-estimator	Depends on $\rho(\cdot)$	Depends on $\rho(\cdot)$ Higher than similar M-estimator.	$\sqrt{n}$ -consistent Asymptotic Normal Very efficient, but less than M-estimator	$O(np^3 f(n, p))$ / iteration	$p < n$
Stahel Donoho Estimator	Yes	1/2	$\sqrt{n}$ -consistent Asymptotic distribution not known	$O(n^2 p)$	None

# Application

---

## EXAMPLE WITH REAL DATASET

### Diabetes Dataset:

1. Five measurement of 145 adult patients.
2. Variables are
  - 2.1 Relative weight.
  - 2.2 Fasting plasma glucose.
  - 2.3 Oral Glucose.
  - 2.4 Insulin Resistance.
  - 2.5 Steady State Plasma Glucose (SSPG).
3. Three groups: Normal, Chemical Diabetic, Obese.
4. Reference: Reaven, G. M. and Miller, R. G. (1979); An attempt to define the nature of chemical diabetes using a multidimensional analysis.

## EXAMPLE WITH REAL DATASET

For univariate samples  $X_1, \dots, X_n$ , one simple measure of outlyingness is the z-score.

$$Z_i = \frac{(X_i - \hat{\mu})}{\hat{\sigma}}, \quad i = 1, 2, \dots, n$$

## EXAMPLE WITH REAL DATASET

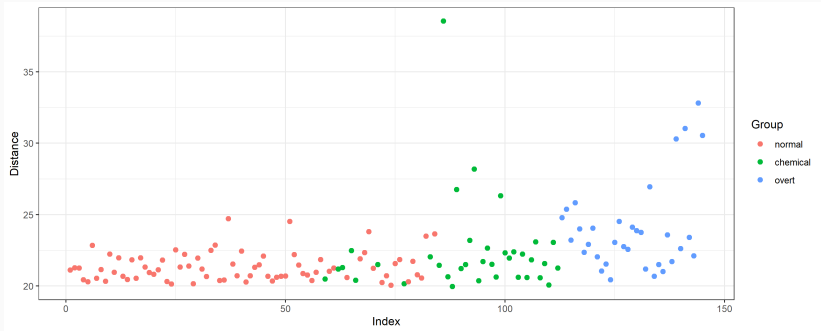
For univariate samples  $X_1, \dots, X_n$ , one simple measure of outlyingness is the  $z$ -score.

$$Z_i = \frac{(X_i - \hat{\mu})}{\hat{\sigma}}, \quad i = 1, 2, \dots, n$$

For multivariate samples  $X_1, \dots, X_n$  with each  $X_i \in \mathbb{R}^p$ , we can look at squared Mahalanobis distance

$$Z_i = (X_i - \hat{\mu})^\top \hat{\Sigma}^{-1} (X_i - \hat{\mu}), \quad i = 1, 2, \dots, n$$

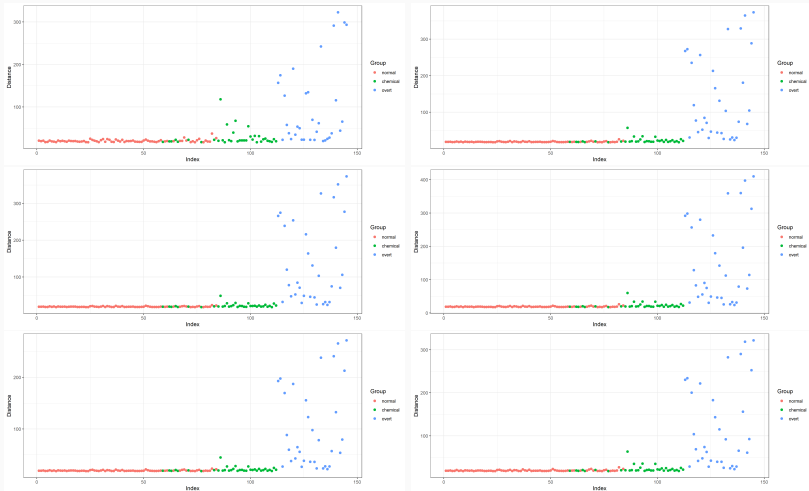
# EXAMPLE WITH REAL DATASET



**Figure 1:** Classical Mahalanobis distance



# EXAMPLE WITH REAL DATASET



**Figure 2:** From Top-Left to Bottom-Right: Mahalanobis distance with W-estimator, MCD, MVE, M-estimator, S-estimator, Stahel Donoho estimator.

# CONCLUSION

1. Robust Statistical Inference can help with the identification of outliers.
2. It is a great way to deal with current high-dimensional datasets without worrying about outliers, where identifying outliers would be challenging.
3. For  $n < p$ , Stahel Donoho estimator should be the first choice as robust location and scatter estimates.
4. For  $n > p$ ,  $M$ -estimators can be a quick robust estimator, whereas MVE and MCD estimators should be used if high breakdown is required.
5. Combining estimators of different data sources to obtain robust estimates, are growing in this era of big data.

**THANK YOU**

---