L20: Noise + Outliers

Mar 31, 2025

Data Mining

Jeff M. Phillips

Types of Noise

Data $X = \{x_1, \dots x_n\}$

- **Measurement Noise** $x_i = \underline{\text{true}(x_i)} + \varepsilon_i$

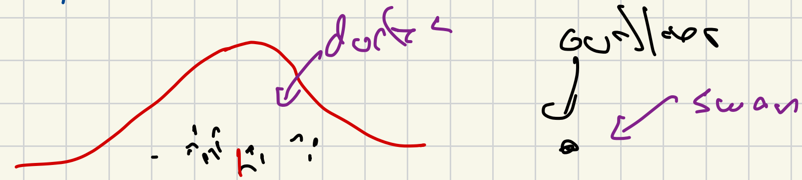
  common Noise $\mathcal{N}(0, I) = \exp\left(\frac{-\|x-\mu\|^2}{2}\right)$ $\varepsilon_i \sim \text{Noise}$

  ↳ loss function (opt - criteria

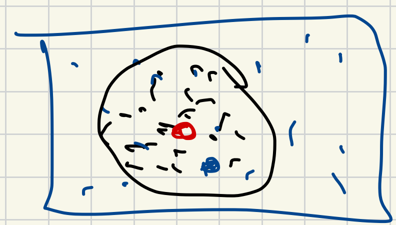
  $\sum_i \|\hat{u} - x_i\|^2$

- **Spurious Readings**

  ↳ outliers



- **Background data**



Dealing w/ Outliers

Goals: 1. Estimate model of inliers

2. Identify outliers, since interesting

Requires assumption inlier data is fit by some model $M \in \mathcal{M}$ ← family of models

Family of Models $\mathcal{M}$ parameters

- think $X \sim N_d(\mu, I)$ $X \subset \mathbb{R}^d$
- k-means clustering / Mix Gaussians
- low-rank + $N(0,1)$ noise → PCA
- linear regression

Outlier Removal

my goal!

1. Fit best model $M^* = \underset{M \in \mathcal{M}}{\arg\max}$ Liklihood $(M; X)$

2. For all $x_i \in X$ calculate e.g. $-\log(\text{Liklihood}(M, x_i))$

   residual $r_i = D(x_i, M)$ k-means

   e.g. $M(x_i) = \phi_S(x_i)$ ← closest cluster center

   $D(x_i, M(x_i)) = \| x_i - \phi_S(x_i) \|$

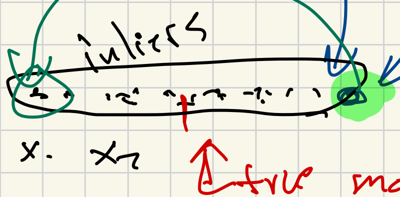
3. Remove $x_i \in X$ s.t. r_i is too big. ?

   needs threshold: user defined

4. Repeat (Goto 1) on $X \setminus \{\text{outliers}\}$.

Mean → $X \subset \mathbb{R}$

too big?



sample mean

$\hat{\mu} \Rightarrow M^* = N(\hat{\mu}, I)$

Goal!

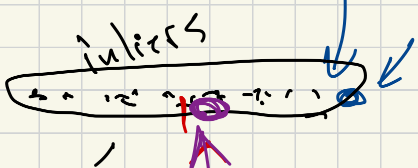
outlier

x_i

residual r_i is very large

Repeat after filtering outliers

$X_1 = X \setminus \{\text{outliers}\}$



$\hat{\mu}_1 \Rightarrow M_1^* = N(\hat{\mu}_1, I)$

Might not work if M^* is too poor an estimate

Robust Estimators $X \subset \mathbb{R}$

Median = sort data, take point at 50%.

has breakdown point of 0.5

point which minimizes

$$\bar{u} = \arg\min_{u \in \mathbb{R}} \sum_{x_i \in X} |x_i - u|$$

---

breakdown point = # points needed to move "outliers" from inliers so $\bar{u}$ outside inliers

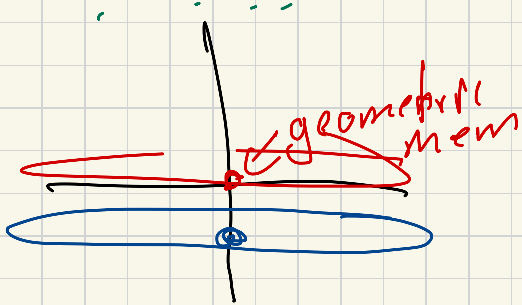
How to compute median in $\mathbb{R}^d$, $d>1$?

as any robust estimator in $\mathbb{R}^d$.

- geometric mean

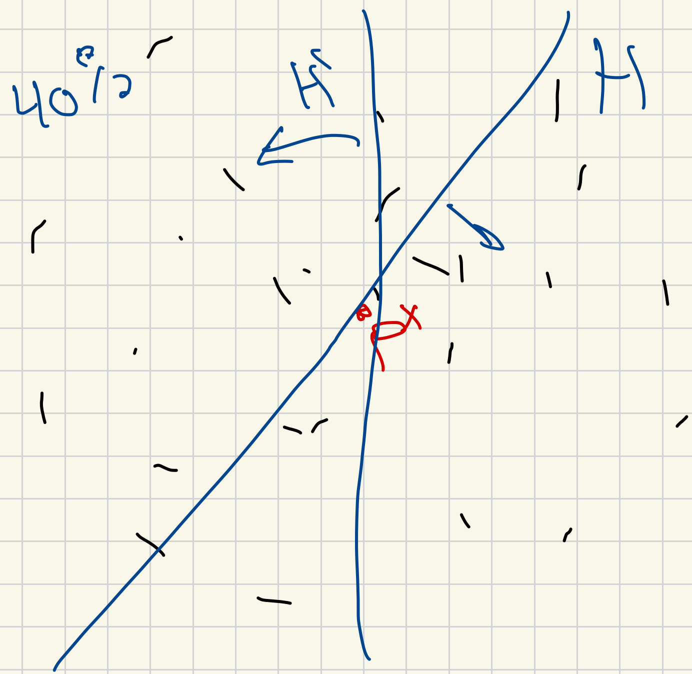
$$\tilde{\mu} = \underset{\mu \in \mathbb{R}^d}{\arg\min} \sum_{x_i \in X} \| x_i - \mu \|$$

has a large breakdown point



* Tukey Median

point $p^* = \max_{p} \min_{H} \left| \text{halfspace } H \text{ contains } (p) \cap X \right|$



gives good estimator

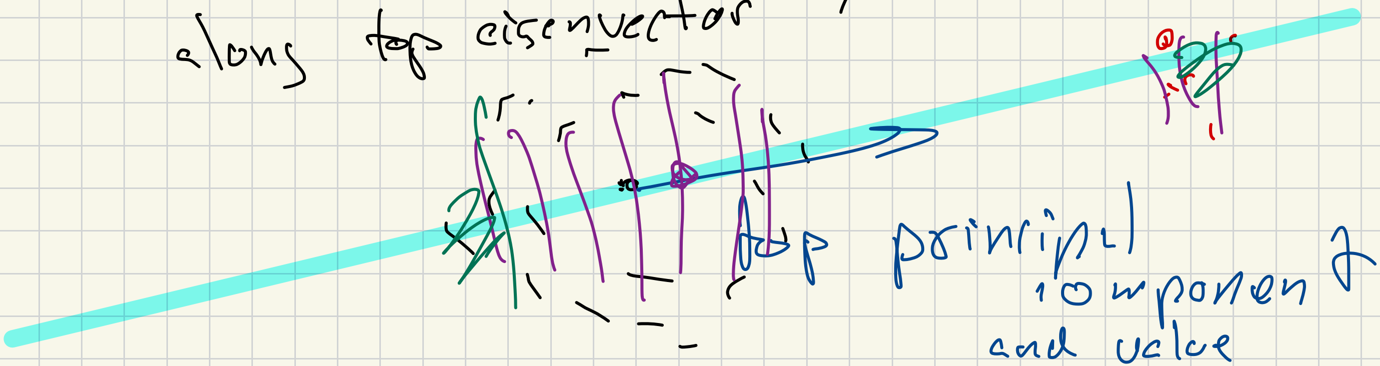
but hard to compute large d

$\Theta(n^d)$

Robust Mean Estimator

$X \sim N(\mu, I)$

1. Estimate sample mean $\hat{\mu} = \text{mean}(X)$
2. Estimate sample covariance $\hat{\Sigma}_1 = \frac{1}{|X|} \sum_{x_i \in X} (x_i - \hat{\mu})(x_i - \hat{\mu})^T$
3. If $\|\hat{\Sigma}_1\|_2$ too big, filter points along top eigenvector



Median of Means

0. Choose parameter k (e.g. $k=5$)

1. Randomly split X into k equal sets $X_1, X_2, \ldots, X_k$

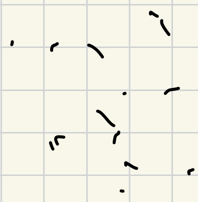
2. Calculate mean of each $X_i \rightarrow \text{mean}(X_i) =: \mu_i$

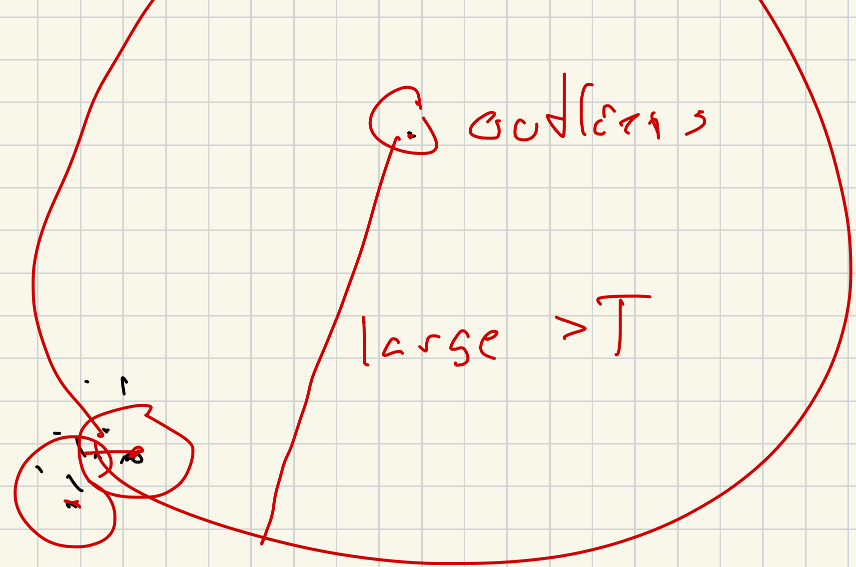
3. Coordinate-wise median of $\{\mu_1, \mu_2, \ldots, \mu_k\}$

Density-based Outliers

When <u>no</u> parametric model [illegible]

Still need good distance D.





- DB Scan (clustering)
  all points not in cluster $\Rightarrow$ outliers

- Parameter k (# points to define neighborhood)
  threshold T
  For each x_i distance D_k to kth closest neighbor. if $> T \Rightarrow$ outlier

- Kernel Density Estimate

$$K(x_i, g) = \exp(-D(x_i, g)^2)$$

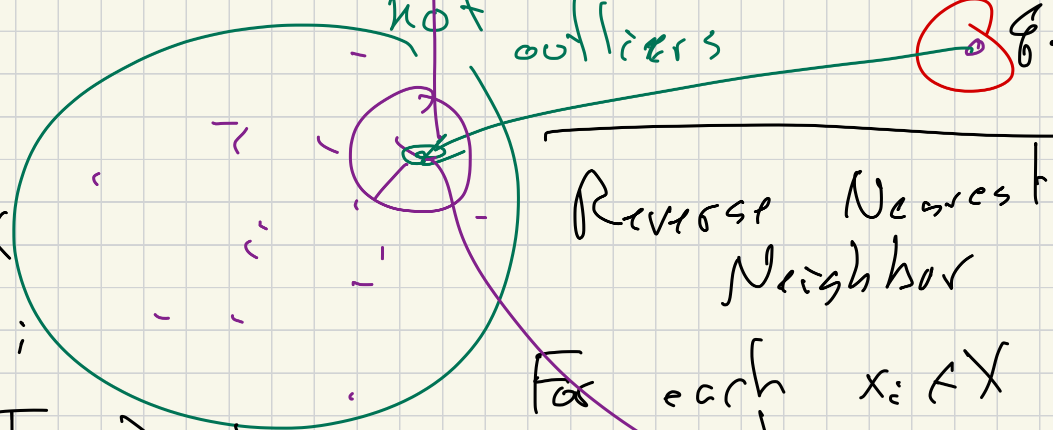
$$KDE_X(g) = \frac{1}{|X|} \sum_{x_i \in X} K(x_i, g)$$

if $\leq T \Rightarrow g$ outlier

All assume density is uniform in inliers,

Let $N(q_i) = x_i \in X$
Nearest to q_i

if $\dfrac{D(q_i)}{D_i(x_i)} > T \Rightarrow$ outlier



Reverse Nearest Neighbor

For each $x_i \in X$
compute

$$D_i = \underset{x_j \in X \setminus x_i}{\arg\min} \|x_i - x_j\|$$

Outlier