Statistical Methods IV Presentation:

Multivariate Median

Sarthak Sarkar (BS2116)

Kundar Dutta (BS2117)

Purushottam Saha (BS2119)

Alaap Kumar Mukhopadhyay (BS2125)

Sattick Das (BS2046)

B.Stat 2nd Year

Indian Statistical Institute, Kolkata

Table of Contents

1. Introduction
2. Coordinate-wise Median
3. Geometric Median
4. Tukey's Median
5. Oja's Median
6. Simplicial Median
7. Convex Hull
8. Application of Multivariate Median : Median based PCA

Table of Contents

1. Introduction
2. Coordinate-wise Median
3. Geometric Median
4. Tukey's Median
5. Oja's Median
6. Simplicial Median
7. Convex Hull
8. Application of Multivariate Median : Median based PCA

Introduction

- For some given **Real numbers** the median is a point estimator of central tendency which has the property that the number of data points that are left to it is same as the number of the data points are right to it.
- For a given set of numbers $x_1, x_2, \cdots x_n$ the median is $\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} |x_i - \mu|$
- If n is odd then median is $x_{(\frac{n+1}{2})}$

  and if n is even then the median is defined to be $\frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$ for uniqueness.
- It is very robust. with respect to the mean.

Breakdown Point

- The replacement breakdown point (RBP) of a location estimator T at the given sample X_n is defined as
  $\text{RBP}(\text{T}, X_n) = \min\left\{ \frac{m}{n} : \sup_{X_n^m} \| T(X_n^m) - T(X_n) \| = \infty \right\}$
  where X_n^m denotes a contaminated sample from X_n by replacing m points of X_n with an arbitrary m points.
- The Breakdown point of a Estimator T is $\lim_{n\to\infty} \max_{X_n} RBP(T, X_n)$
- The Replacement Breakdown point of median is $\frac{\lfloor n+1 \rfloor}{2n}$ and mean is $\frac{1}{n}$
- Asymptotically we see that the Breakdown point of median is 50% but for mean that is 0

Table of Contents

1. Introduction
2. Coordinate-wise Median
3. Geometric Median
4. Tukey's Median
5. Oja's Median
6. Simplicial Median
7. Convex Hull
8. Application of Multivariate Median : Median based PCA

Coordinate-wise Median

- For multi dimensional data **Coordinate-wise Median** is just taking the median of the each components(or dimension) of the data and joining them to get the median of the data points.

$$\hat{\theta} = [\hat{\theta_1}, \hat{\theta_2}, \cdots, \hat{\theta_p}]$$

  where $\hat{\theta_i}$ is the median of the i th dimention of the data points.
- It has breakdown point of 50 % because each of the coordinate is the median of some real numbers and each of them has breakdown point of 50 %.
- It is not equivariant with respect to Euclidean similarity transformations such as rotations.

Table of Contents

1. Introduction
2. Coordinate-wise Median
3. Geometric Median
4. Tukey's Median
5. Oja's Median
6. Simplicial Median
7. Convex Hull
8. Application of Multivariate Median : Median based PCA

Geometric Median

- In one dimension we find the $\arg\min_{\mu} \sum_{i=1}^{n} |x_i - \mu|$ we generalize the idea for $x_i \in \mathbb{R}^p$
- For n points in $\mathbb{R}^p$ we define their Geometric median or Spatial median to be

$$\hat{\mu} = \arg\min_{\mu \in \mathbb{R}^p} \sum_{i=1}^{n} \|x_i - \mu\|_2$$

  where $\|.\|_2$ is the euclidean norm.
- It is equivariant with respect to Euclidean similarity transformations such as translations and rotations.
- It has a High Breakdown point of 50% .

Geometric Median

- Since the objective function that we are minimizing is convex the objective function has less value in the convex hull of the data points compared to outside of the data points and since the function is continuous on the convex hull the minima is achieved.
- But it might not be unique, if no three points are collinear then the geometric median is unique.

Geometric Median

- For finding the Geometric median we us a form of iteratively re-weighted least squares algorithm called Weiszfeld's algorithm
- If we differentiate the objective function w.r.t. μ which is differentiable for all $\mu \in \mathbb{R}^p$ except the data points $x_1, x_2, \cdots x_n$ we get,

$$Df(\mu) = \sum_{i=1}^{n} \frac{x_i - \mu}{\|x_i - \mu\|_2}$$

or for minima we equate this to 0 to get

$$\mu = \frac{\sum_{i=1}^{n} \frac{x_i}{\|x_i - \mu\|_2}}{\sum_{i=1}^{n} \frac{1}{\|x_i - \mu\|_2}}$$

- For solving the above equation we use Weiszfeld's algorithm.

Geometric Median

- For finding the median of $x_1, x_2, \cdots, x_n \in \mathbb{R}^p$ first we choose a $\mu_0 \in \mathbb{R}^p$ not among the data points, and define

$$F(\mu) = \frac{\sum_{i=1}^{n} \frac{x_i}{\|x_i - \mu\|_2}}{\sum_{i=1}^{n} \frac{1}{\|x_i - \mu\|_2}}$$

- Define $\mu_k = F(\mu_{k-1}) = \frac{\sum_{i=1}^{n} \frac{x_i}{\|x_i - \mu_{k-1}\|_2}}{\sum_{i=1}^{n} \frac{1}{\|x_i - \mu_{k-1}\|_2}}$
- Continue the iteration until converges. If for some μ_k, $\mu_k \in \{x_1, x_2, \cdot, x_n\}$ just ignore that x_i in the next iteration.

Table of Contents

1. Introduction
2. Coordinate-wise Median
3. Geometric Median
4. Tukey's Median
5. Oja's Median
6. Simplicial Median
7. Convex Hull
8. Application of Multivariate Median : Median based PCA

Tukey's Median

- In Tukey's median we generalize the concept of the depth of a point. In $\mathbb{R}$ the median is the point with the highest depth.
- Consider n points $x_1, x_2, \cdots, x_n$ in $\mathbb{R}^p$. Now for a point $x \in \mathbb{R}^p$, define the Half-Space Depth of the point as the minimum number of points contained in a half-space containing that point.

$$HD(x) = \min_{H:x\in H} \sum_{i=1}^{n} \mathbb{1}_{[x_i \in H]}$$

- Tukey's median is any point with maximum possible depth.

$$\hat{\theta} = \max_{x\in\mathbb{R}^p} HD(x)$$

- The Additive Breakdown point of Tukey's median is $\frac{1}{p+1}$ where p is the dimention of the data points.

Algorithm

Table of Contents

1. Introduction
2. Coordinate-wise Median
3. Geometric Median
4. Tukey's Median
5. Oja's Median
6. Simplicial Median
7. Convex Hull
8. Application of Multivariate Median : Median based PCA

Oja's Median

- In Oja's median we generalize the concept of the simplex volume. In $\mathbb{R}$ the median is the point which minimises the volume of the simplexes that are made with p points from dataset and the median.
- Let us define $Vol(S[x_1, x_2, ..., x_p, x_{p+1}])$ as the volume of p -dimensional simplex $S[x_1, ..., x_p, x_{p+1}]$ with vertices $x_1, ..., x_{p+1}$. Now we define the Oja's median θ as

$$\hat{\theta} = \arg\min_{x} \sum_{i_1 \leq i_2 \leq .... \leq i_n} Vol(S[Xi_1, Xi_2, ..., Xi_p; x])$$

  where $X_1, \cdots, X_n$ is a given data set and $i_1, ..., i_p$ are distinct p integers from 1,...,n.
- Oja's median is generally not unique.
- It has very low breakdown point. In $\mathbb{R}^2$ it is $\frac{2}{n+2}$.
- It has breakdown point 0 iff the datapoints lie on a (p-1) dimensional subspace of $\mathbb{R}^p$

Table of Contents

1. Introduction
2. Coordinate-wise Median
3. Geometric Median
4. Tukey's Median
5. Oja's Median
6. Simplicial Median
7. Convex Hull
8. Application of Multivariate Median : Median based PCA

Simplicial Median

- In Simplicial median we generalize the concept of the simplex depth of a point. In $\mathbb{R}$ the median is the point which is on the maximum number of simplexes that are made with (p+1) points from dataset.
- A p-dimensional simplex is a convex hull of p+1 points in $\mathbb{R}^p$. Consider n points in $\mathbb{R}^p$. For any p+1 points out of them, consider the corresponding simplexes. A point that belongs in the most number of simplexes, is called the Simplicial Median for the data.
- It has a Additive breakdown point of less than $\frac{1}{p+2}$

Table of Contents

1. Introduction
2. Coordinate-wise Median
3. Geometric Median
4. Tukey's Median
5. Oja's Median
6. Simplicial Median
7. Convex Hull
8. Application of Multivariate Median : Median based PCA

Convex Hull

- We generalise the idea of data peeling into $\mathbb{R}^p$. In $\mathbb{R}$ the median is the point which remains by removing the outer points of the data set.
- Consider n points in $\mathbb{R}^p$. Consider the convex hull of those points. If removing the perimeter of the convex hull still results into existance of points, continue this process till possible, and hence we get the deepest (in a sense) convex hull. All the points on the Convex Hull defined as the Median .
- For some cases it has a breakdown point of $\frac{1}{n}$

Table of Contents

1. Introduction
2. Coordinate-wise Median
3. Geometric Median
4. Tukey's Median
5. Oja's Median
6. Simplicial Median
7. Convex Hull
8. Application of Multivariate Median : Median based PCA

Robust PCA

PCA is one of the most famous dimension reduction algorithms, but it uses the mean function, which is of course, not robust. Hence for more corrupted data, PCA does not give a satisfactory result. In search for the robust wayout, we use Multivariate Median to create the Median Correlation Matrix, and apply the similar process to the same, to get a more robust PCA. We use this technique for identifying moving objects in surveillance cameras.

Approach/ Algorithm

The main idea in Median based PCA is to change the central tendency while shifting the origin, from mean to median, or change the Covariance Matrix to Median Covariance Matrix.
So we define Median Covariance Matrix by :

$$\tilde{\Sigma} = median\{x_i{}^T x_i\}_i$$

where x_i is the i'th observation.
After that, just as we find for PCA, the direction with the most variance on projection distances, we do the same for Median Covariance Matrix in place for Covariance Matrix, and we get our Robust PCA.
The Median being more robust estimator for central tendency, Robust PCA gives better results over PCA if the sample has outliers, as mean, as a measure of central tendency is very prone to outliers.
Also, whenever the sample is corrupted with some error, it is always a good idea to consider Robust PCA over PCA, as the former using medians, have significant higher breakdown point. So for some pixels being corrupt for a video data from Surveillance Camera, it is better to

Facilities and Implementation

We have wrote the Python program to find the Median Based Algorithm.