# Subspace rotations for high-dimensional outlier detection

2021311169 서현기

## Introduction

- It is difficult to detect outliers in HDLSS(High Dimension Low Sample Size) setting
  - Insufficient observations to characterize the 'regular' behavior
  - Observations become deterministic converging to th vertices of a simplex

    $\rightarrow$ It makes impossible to take  ordinary 'distance' as a measure of abnormality
  - Because most of test depends on large sample approximation, it is chiallenging to construct a valid hypothesis test in testing abnormal observations due to small sample size and high dimensionality.
- Existing Methods detecting outliers in HDLSS settings and their limitation.
  - Comedian distance - computationally expensive because it requires inversion of $d \times d$ matrix
  - Minimum covariance determinant estimator - may not work in real world because of their strong assumption
  - kurtosis of the PC score as a measure for abnormality - PC directions are not generally consistent with increasing dimensionality
  - Distance-based outlier detection method - produce meny false negatives when outliers are clusterd
  - Angle-based outlier detection method - not equipped with a formal testing procedures.

## Preliminaries

- Notations
  - $\mathcal{O}_N$ : orthogonal group of order N
  - $\mathcal{V}_{m,N}$: stiefel manifold s.t $\mathcal{V}_{m,N}\{V \in \mathbb{R} | V'V = I_m\}$

    $\rightarrow \mathcal{V}_{N,N} = \mathcal{O}_N$
- Left-spherical distribution

  Let $\mathbf{X}$ be an $N \times d$ random matrix according to a probability distribution $\mathbb{P}$. if $O\mathbf{X}$ is identically distributed as $\mathbf{X}$ , for all $O \in \mathcal{O}_N$ , then $\mathbb{P}$ is called a left-spherical distribution, denoted by $\mathbb{P} \in LS_{N,d}$
  - Which means a distribution symmetric for all axes and rotations
  - necessary conditions are zero mean and identity row-wise covariance matrix(uncorrelated obs)
  - Location-shifted left-spherical family :  $\mathbf{X} - E(\mathbf{X})$ is left-spherical distribution
  - Example : Matrix normal, Matrix T, a scale-mixture of left-spherical distributions
- Randomization Test : Testing methods that regenerate data from same distribution and Test the hypothesis through the empirical CDF and the original data.

$\rightarrow$ Trough the property of LS distribution, We are going to regenerate data by multiplying rotation matrix ($RX$) and this can be viewed as the new data which is generated from the same distribution with $X$. with these new data, randomization test is conducted.

## Subspace rotations

- Assumption : $E(X)$ is known **up to its column space** and can be written as $E(X) = M_0 B'$ where $M_0$ is $N \times m_0$ basis vector and orthonormal , $B$ is unknown coefficient matrix.
- Let $LS_{N,d}(M_0)$ be the set of location-shifted left-spheical distributions with mean matrix whose column space is spanned by $M_0$
- Consider the following subgroup of $\mathcal{O}_N$ :

  $$\mathcal{R}(M_0) = \{R|R = M_0 M_0' + M_1 O M_1, O \in \mathcal{O}_{m1}\}$$

  where $M_1$ is an $N \times m_1$ matrix whose columns constitute an orthonormal basis of $\mathrm{span}(M_0)^{\perp}$
- Theorem

  let $X \sim \mathbb{P} \in LS_{N,d}(M_0)$ and $R \sim \mathcal{U}(\mathcal{R})$, where $X \text{ and } R$ are independent. Then, $RX$ and $X$ are identically distributed and $RX$ and $R$ are independent
- By using the theorem above, we can make randomization test.

  $$H_0 : \mathbb{P} \in LS_{N,d}(M_0)$$
  $$F_{t|X}(z|X) = \int_{\mathcal{R}} 1\{t(RX) \leq z\}[dR]$$

  where $t(RX)$ is chosen test statistic such that the rejection region has the form of $t(X) > c$
- The conditional distribution above is an unbiased estimator of the true distribution $F_t(z) = Pr(t(X) \leq z)$

## Application to high-dimensional outlier detection

### measure of abnormality

- We use Distance to Hyper plane(DH) as a measure of abnormality which means **the closest $L_2$ distance from $x_i$ to $Aff(X_s)$** where $X_s$ is a $n_s \times d$ row-wise sub-matrix of $X_{-i}$

  $$DH(x_i|X_s) = \|(I_d - P_s)(x_i - \bar{x_s})\|_2$$

  where $P_s$ is the projection matrix onto the rowspace of $X_s$, $\bar{x}_s$ is the average of the rows of $X_s$

### Screening of candidate outlier

- At first, We must choose $n^*_{out}$ (number of potential outlier) and $s$ (number of regular points that are identified by median pairwise distance)
- $n^*_{out}, s \leq \lfloor N/2 \rceil$
- Calculate pairwise distance between all points of $X$
- Find median distance $\xi_i$ for each observation

- Rearrange $x_1, \cdots, x_N$ so that $\xi_1 \leq \cdots \leq \xi_N$
- Set $\mathcal{S} = \{x_1, . \cdots, x_s\}$ and $\mathcal{X} - \mathcal{S} = \{x_{s+1}, \cdots, x_N\}$
- Calculate $DH$ for all obs in $\mathcal{X} - \mathcal{S}$ from $\mathcal{S}$

$$D_k = DH(x_k|X_{\mathcal{S}})$$

- Rearrange $x_{s_1}, \cdots, x_N$, so that $D_{s+1} \leq \cdots \leq D_N$
- Set $\mathcal{X}_0 = S \cup \{x_{s+1}, \cdots, x_{N-n^*_{out}}\}$, and $\mathcal{X}_1 = \mathcal{X} - \mathcal{X}_0$

### Sequential SR tests on candidate outliers

- let $Y_j = [X'_0, x^*_j]$ where $x^*_j \in \mathcal{X}_1$. Thus $Y_j$ is $n \times d$ matrix where $n = n^*_{in} + 1$
- $x^*_j$ means $j$th point in $\mathcal{X}_1$, rearranged by $t_1 \geq \cdots \geq t_{n^*_{out}}$ where $t_j$ is $DH(x^*_j|X_0)$
- Carry out sequential SR tests

  $H_{0,j} : Y_j \sim \mathbb{P} \in LS_{n,d}(J_{n,1})]$

  $\rightarrow$ The idea of this test : if potential outlier is a regular data, $Y_j$ is still Left-spherical data (because they are from same distribution)
- In SR tests, regenerate data by $Y^{(k)}_j = R_k Y_j$

  $R_k \overset{iid}{\sim} \mathcal{U}(\mathcal{R})$, $k \in \{1, \cdots, K\}$ where $\mathcal{R} = \mathcal{R}(n^{-1/2}J_{n,1})$
- Save the most far $DH$ of each rotation, denoted by $t^k_j$
- Set critical value

  $\hat{c}_{\alpha,j} = min\{z|\hat{F}_{t|Y_j}(z|Y_j) \geq 1 - \alpha\},\ \text{where}\ \hat{F}_{t|Y_j}(z|Y_j) = K^{-1}\sum 1(t^{(k)}_j \leq z)$
- If $t_j \geq \hat{c}_{\alpha,j}$, declare $x^*_j$ as an outlier.
- Starting with $x^*_1$ , repeat the above tests until we fail to reject $H_{0,j}, j \leq n^*_{out}$.

  At this point, $x^*_1, \cdots, x^*_{j-1}$ are identified as outlier

### Computational complexity

- Although $DH$ is calculated as $DH(y^{(k)}_i|Y_{j,-i}) = \|(I_d - P^{(k)}_{-i})(y^{(k)}_i - \bar{y}^{(k)}_{-i})\|_2$,

  It can be replaced by $DH(y^{(k)}_i|Y_{j,-i}) = \frac{2}{\|Y^+_{c,j}\ \underline{l_i}\|_2}$

  where $Y^+_{c,j}$ is the Moore -Penrose inverse of $Y_{c,1}$, $\underline{l_i}$ is label vector with $l_{j,i} = 1$if $j = i$ and 0 otherwise
- By this replacement, The dimension of Projection matrix becomes $(N-1) \times (N-1)$ which is far smaller then $d \times d$ in HDLSS setting.

## Asymptotical properties

- Asymptotic : $N$ is fixed, $d \rightarrow \infty$
- Conditions:
  1. the fourth moments of the entries of the data vectors are uniformly bounded
  2. $\sum^d_{l=1}\{E(x_l) - E(x^0_l)\}^2/d \rightarrow \mu^2$ (location differences are converge)

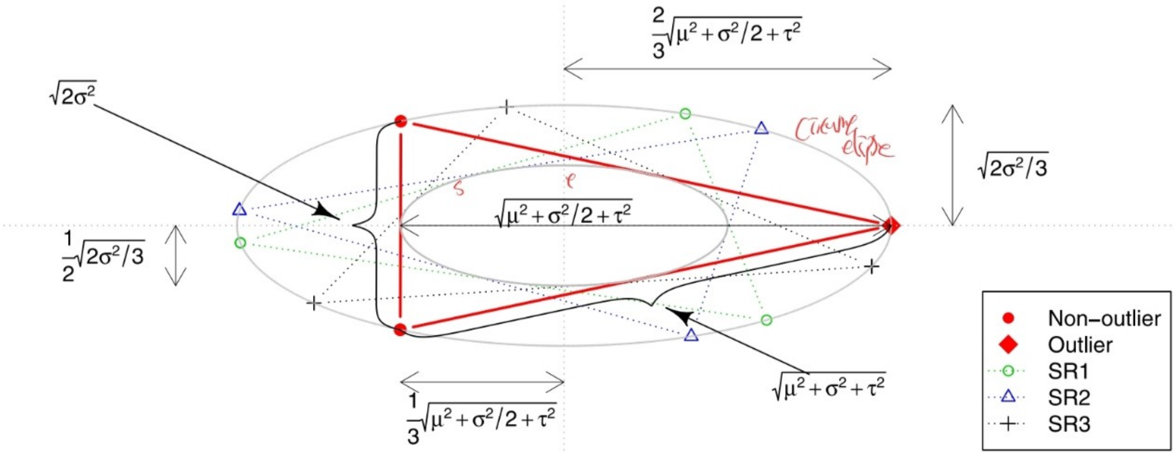
3. $\sum_{l=1}^{d} var(x_l)/d \to \sigma^2$ ( variations of variables in regular data are converge)
4. $\sum_{l=1}^{d} var(x_l^o)/d \to \tau^2$ ( variations of variables in regular data are converge)
5. for both $x$ and $x^o$, there exists a permutation of entries such that the sequence of the variables are $\rho-$mixing for functions that are dominated by quadratics. (mild condition to achieve the law of large numbers)
6. True number of outliers and non-outliers satisfy $n_{out} + 1 < n_{in}$

- Assume Conditions above is hold. If $\mu^2 + \tau^2 > \sigma^2$ then

$$\lim_{d \to \infty} Pr(\mathcal{S} \cap \mathcal{J} \neq \emptyset) = 1$$
$$\lim_{d \to \infty} Pr(\mathcal{J} \subset \mathcal{X}_1) = 1$$

where $\mathcal{J}$ is the set of true outliers.

### Geometric interpretation of the test




- As $d \to \infty$, The rotated data makes a simplex sharing common set of Steiner inellipse and circumellipse with original data.
- Subspace rotation test would compare the maximum height of the original data simplex(triangle with red line) with the maximum height of other triangles(triangle with dotted line).
- If the potential outlier is not the true outlier(generate from a same distribution with regular data), the simplex have same side lengths (such as Equilateral triangle). Thus, the height of original data is lower than the others.
- But if the potential outlier is true outlier, the simplex does not have same side lengths and is likely to have the longest height

## Compare with other methods

- Comparing SR test with Comedian(COM), Distance based method (DSO), PCout(PCO) and MDP
- In three different settings
  1. Auto regressive (Adjacent variables are correlated)
  2. Compound symmetry (All variables are equally correlated)
  3. Geometric Decay
- In Simulating results and real data analysis,

  All methods have good performance of True positive rates.

- Other methods have bad performance of False positive rates in some settings, but SR test method have  high performance of False positive rates