

Detecting Outliers in HDLSS Data

Jeongyoun Ahn

Korea Advanced Institute of Science and Technology

CDAR @ Berkely 2022.

Joint work with Hee Cheol Chung

Outliers in HDLSS data

- High-throughput data are more likely to have abnormal observations.
- Challenges in developing methodology
 - Hard to distinguish outliers from non-outliers as d increases.
 - Distance measure when $n \ll d$.
 - Significance test.
- There are only a few existing methodologies.
 - Rely on n -asymptotic test.
 - Use a high-dimensional covariance estimator.
 - Tend to have larger false positives.

Outlier detection for HDLSS data

When $N \ll d$

- PCont (PCO) by Filzmoser et al. (2008): Kurtosis based approach on the PC scores.
- CoMedian (COM) by Sajesh and Srinivasan (2012): Based on robust estimates of mean and covariance matrix.
- MDP by Ro et al. (2015): Only diagonal elements of covariance matrix are considered.
- DSO by Ahn et al. (2020): Sequential elimination based on LOO distance.

Outlier detection for HDLSS data

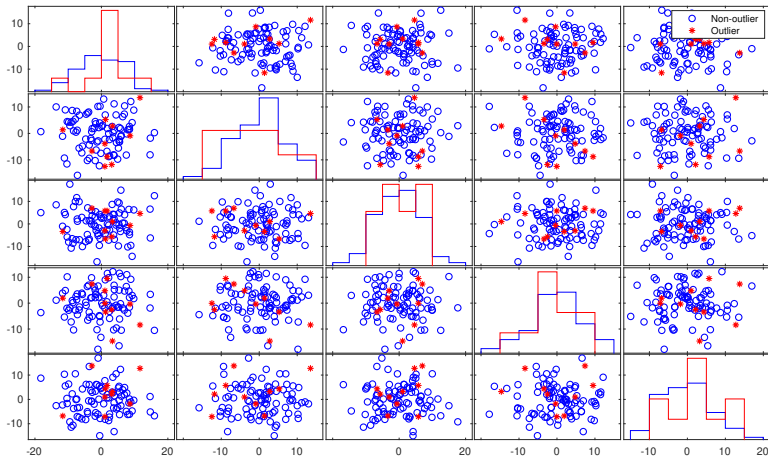
When $N \ll d$

- PCout (PCO) by Filzmoser et al. (2008): Kurtosis based approach on the PC scores.
- CoMedian (COM) by Sajesh and Srinivasan (2012): Based on robust estimates of mean and covariance matrix.
- MDP by Ro et al. (2015): Only diagonal elements of covariance matrix are considered.
- DSO by Ahn et al. (2020): Sequential elimination based on LOO distance.

All rely on large sample approximation in testing respective abnormality measures.

Outliers in HDLSS data

- $\mathbf{x}_i \stackrel{iid}{\sim} N(0, \Sigma)$, where $i = 1, \dots, 90$, $\Sigma = 0.8^{|l-l'|}$ and $1 \leq l, l' \leq 2000$.
- $\mathbf{x}_j^o \stackrel{iid}{\sim} N(10\mathbf{u}_j, \Sigma)$, where $j = 1, \dots, 10$ and $\mathbf{u}_j \stackrel{iid}{\sim} \text{Unif}(\mathcal{V}_{1,d})$.



The Proposed Method

Two Stage Procedure

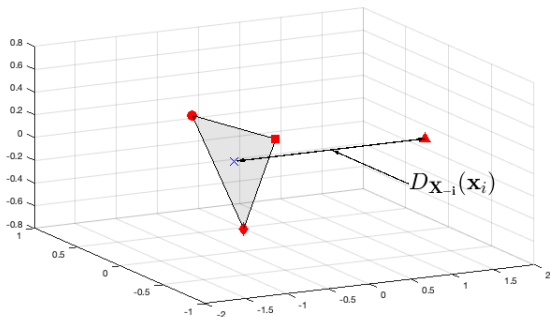
1. Identify “surely” non-outliers and candidate outliers.
 - New outlyingness measure for HDLSS
 - Sure screening property

2. New nonparametric test for candidate outliers.
 - Individual outliers are tested against the surely non-outliers.
 - Random rotation to generate null distribution.
 - HDLSS asymptotic power

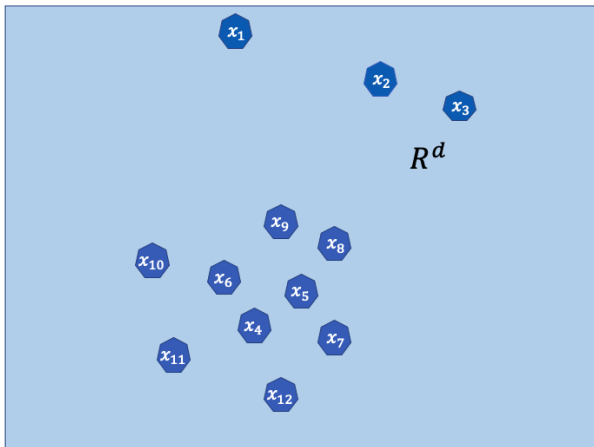
Measure of Outlyingness

- $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^T$, where $\mathbf{x} \in \mathbb{R}^d$.
- Let \mathbf{x}_i^T be the i th row of \mathbf{X} and \mathbf{X}_{-i} be a row-wise sub-matrix of \mathbf{X} without \mathbf{x}_i .
- Let \mathbf{P}_{-i} be the projection matrix onto the row space of \mathbf{X}_{-i} and $\bar{\mathbf{x}}_{-i}$ be the mean of \mathbf{X}_{-i} .
- We use Distance to Hyperplane (DH):

$$D_{\mathbf{X}_{-i}}(\mathbf{x}_i) = \|(I_d - \mathbf{P}_{-i})(\mathbf{x}_i - \bar{\mathbf{x}}_{-i})\|_2.$$

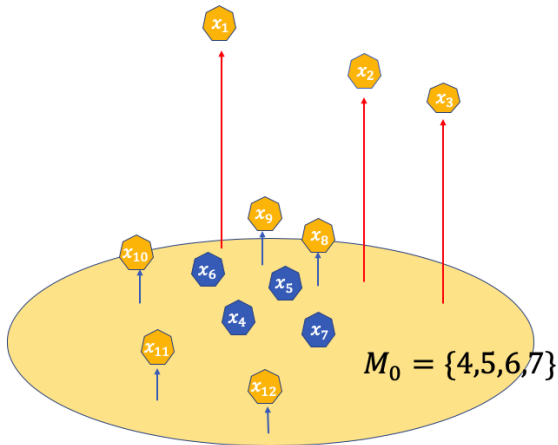


Stage I: Screening Candidate Outliers (1/4)



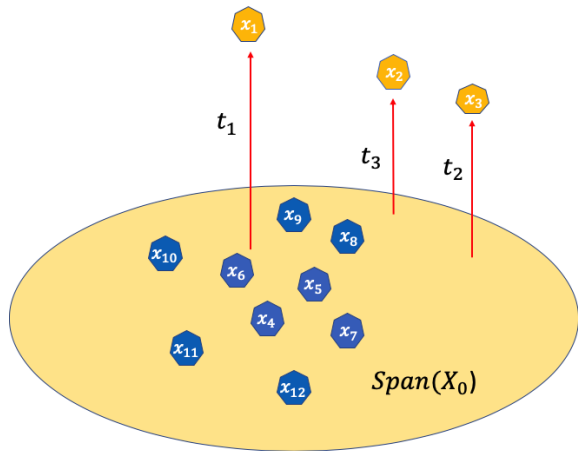
Stage I: Screening Candidate Outliers (2/4)

Choose “inliers” based on median pairwise distance.



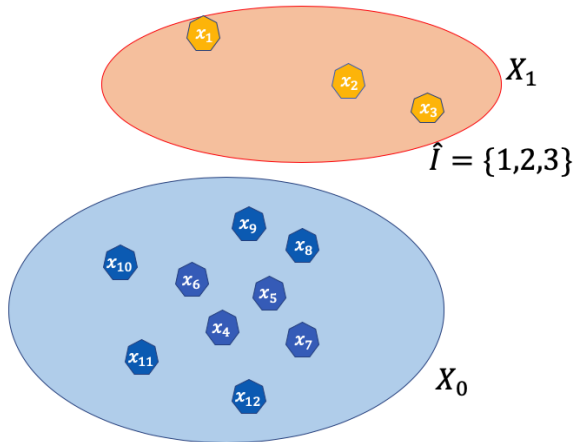
Stage I: Screening Candidate Outliers (3/4)

Candidate outliers based on DH to “inliers”



Stage I: Screening Candidate Outliers (4/4)

"Candidate outliers" vs. "Surely non-outliers".



Stage II: Testing Outliers

For each $\mathbf{x}_i \in \mathcal{X}_1$,

- Test whether \mathbf{x}_i is **too far** from \mathcal{X}_0 .
- Test statistic: DH distance

$$D_0(\mathbf{x}_i) = \|(I_d - \mathbf{P}_0)(\mathbf{x}_i - \bar{\mathbf{x}}_0)\|_2.$$

- Q: How to obtain a null distribution?

Stage II: Testing Outliers

For each $\mathbf{x}_i \in \mathcal{X}_1$,

- Test whether \mathbf{x}_i is **too far** from \mathcal{X}_0 .
- Test statistic: DH distance

$$D_0(\mathbf{x}_i) = \|(I_d - \mathbf{P}_0)(\mathbf{x}_i - \bar{\mathbf{x}}_0)\|_2.$$

- Q: How to obtain a null distribution?
- **Random Rotations**

Background: Multivariate Normal

From *Dempster (1969)*

- X : an $N \times d$ data matrix from $N_d(0, \Sigma)$.
- QR decomposition:

$$X = X_Q X_R$$

- X_Q is uniformly distributed in a Stiefel manifold $\mathcal{V}_{n,r}$, $r = \text{rank of } X$.
- X_R is a sufficient statistic for Σ .
- X_Q and X_R are independent.

Background: Rotation for Multivariate Normal

From *Langsgrud (2005)*

- Generate X_Q^* from $\mathcal{V}_{n,r}$.
- Obtain new data conditioning on the sufficient statistic for Σ

$$X^* = X_Q^* X_R$$

- X^* is a **randomly rotated version** of X .
- $X^* \sim N_d(0, \Sigma)$.

Left-spherical Distribution

Definition

Let \mathbf{X} be an $N \times d$ random matrix according to a probability distribution \mathbb{P}_N . If $R\mathbf{X}$ is identically distributed as \mathbf{X} , for all $R \in \mathcal{O}_N$, then \mathbb{P}_N is called a left-spherical distribution, denoted as $\mathbb{P}_N \in LS_{N,d}$.

Note:

- Matrix normal, matrix t and a scale-mixture of those are included.
- Rows must be uncorrelated.

Random Rotations

$t(\mathbf{X})$: a given statistic

\mathbf{R} : uniformly on \mathcal{O}_N

Rotation Distribution

- For left-spherical \mathbf{X} , $t(\mathbf{R}\mathbf{X})$ is identically distributed as $t(\mathbf{X})$
- Conditioning on \mathbf{X} , the distribution function of $t(\mathbf{R}\mathbf{X})$ as

$$F_{t|\mathbf{X}}(z|\mathbf{X}) = \int_{\mathcal{O}_N} 1\{t(R\mathbf{X}) \leq z\} [dR], \quad (1)$$

where $1(\cdot)$ denotes an indicator function.

Subspace Rotations for Nontrivial Mean

- $\mathbf{X} - E(\mathbf{X}) \sim LS_{N,d}$.
- Or, $\mathbf{X} \sim \mathbb{P}_N \in LS_{N,d}(M)$, where $E(\mathbf{X}) = MB$ and M is a known $N \times m_0$ full rank matrix.
- Orthogonal transformations:

$$\mathcal{Q}_N = \{L_N = MM^T + M_{\perp}RM_{\perp}^T, R \in \mathcal{O}_{N-m_0}\}$$

Theorem 1

Let $\mathbf{X} \sim \mathbb{P}_N \in LS_{N,d}(M)$ and $\mathbf{L}_N \sim Unif(\mathcal{Q}_N)$, where \mathbf{X} and \mathbf{L}_N are independent. Then $\mathbf{L}_N\mathbf{X}$ and \mathbf{L}_N are independent, and $\mathbf{L}_N\mathbf{X}$ and \mathbf{X} are identically distributed.

Subspace Rotation Tests

- $H_0 : \mathbf{X} \sim \mathbb{P}_N \in LS_{N,d}(M)$.
- Under H_0 , $t(\mathbf{X}) \stackrel{d}{=} t(\mathbf{L}_N \mathbf{X})$, where $\mathbf{L}_N \sim \text{Unif}(\mathcal{Q}_N)$.
- Estimate $F_t(z)$ using the SR distribution

$$F_{t|\mathbf{X}}(z|X) = \int_{\mathcal{Q}_N} \mathbf{1}\{t(L_N X) \leq z\} [dL_N], \quad (2)$$

where $[dL_N] = (dL_N) / \text{Vol}(\mathcal{O}_m)$.

Theorem 2

The conditional distribution in (2) is an unbiased estimator of the true distribution function $F_t(z) = \Pr(t(\mathbf{X}) \leq z)$ in the sense that $\mathbb{E}_{\mathbf{X}}\{F_{t|\mathbf{X}}(z|\mathbf{X})\} = F_t(z)$.

Subspace Rotation Tests

Let $c_\alpha(X)$ be the critical value based on the SR distribution.

$$\phi_{SR}(X) = 1\{t(X) \geq c_\alpha(X)\}. \quad (3)$$

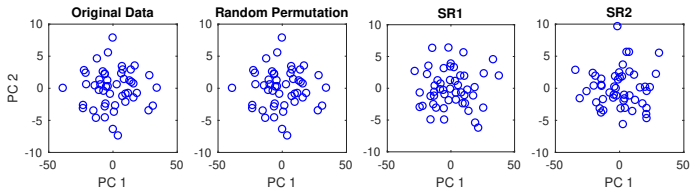
Theorem 3

If the test statistic $t(\cdot)$ is continuous, then the p -value of the SR test is uniformly distributed over $(0, 1)$, and the SR test is a size- α test.

Hypothesis Testing via Random Rotations

Permutations vs Rotations

- It considered as a continuous extension of the permutation tests.
- Random rotations perturb the geometric configuration of the data while preserving its location and volume.



Application: Test of Independence

Multivariate normal with two sets of variables

- $[X, Y] \sim N_{p+q}(0, \Sigma)$
- Partitioned covariance

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}$$

- Consider a hypothesis

$$H_0 : \Sigma_{xy} = 0$$

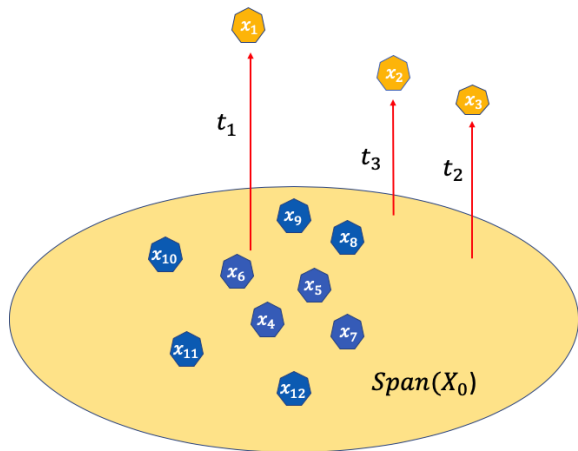
- $g(X, Y)$: a test statistic, such as LRT

Application: Test of Independence

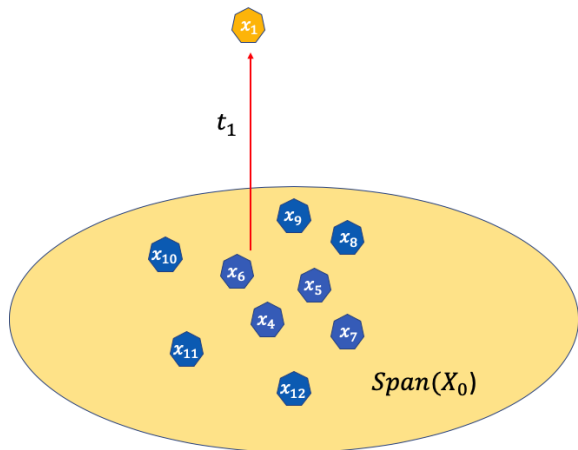
The null distribution $g(X, Y)$ under H_0 can be simulated via rotating data repeatedly.

- $g(R_1 X, R_2 Y)$, where $R_1, R_2 \sim \mathcal{O}_n$
- $g(RX, Y)$ or $g(X, RY)$, where $R \sim \mathcal{O}_n$

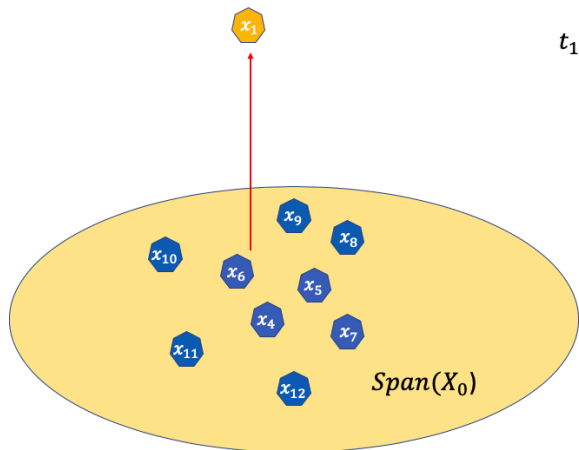
Stage II: Sequential SR Tests on Candidate Outliers (1/8)



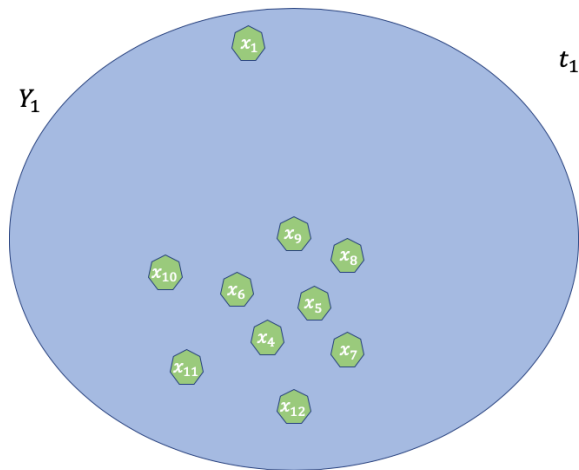
Stage II: Sequential SR Tests on Candidate Outliers (2/8)



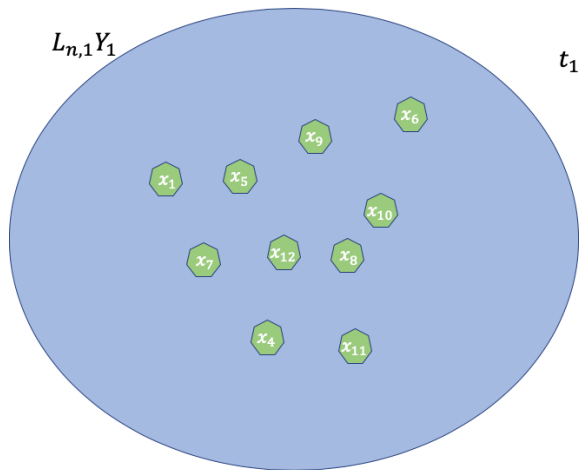
Stage II: Sequential SR Tests on Candidate Outliers (3/8)



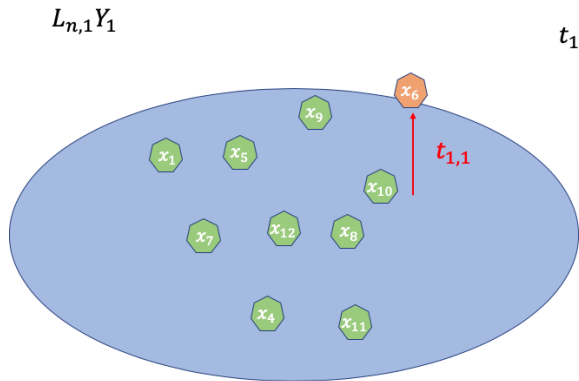
Stage II: Sequential SR Tests on Candidate Outliers (4/8)



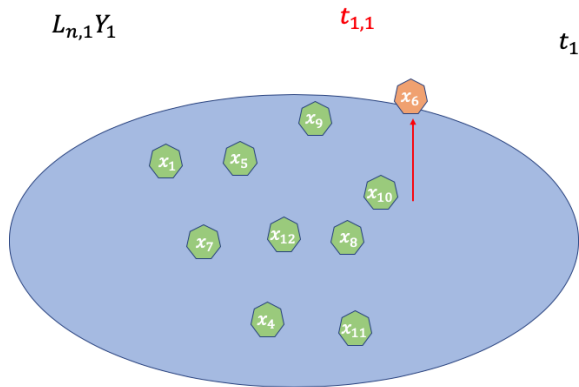
Stage II: Sequential SR Tests on Candidate Outliers (5/8)



Stage II: Sequential SR Tests on Candidate Outliers (6/8)



Stage II: Sequential SR Tests on Candidate Outliers (7/8)

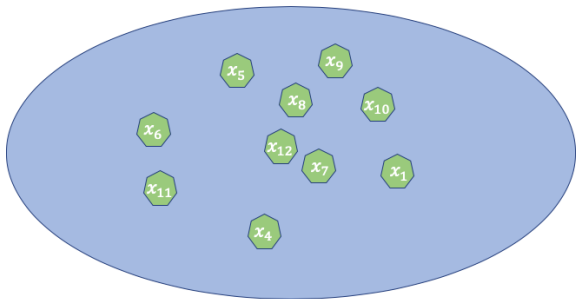


Stage II: Sequential SR Tests on Candidate Outliers (8/8)

$L_{n,K}Y_1$

$\{t_{1,1}, t_{1,2}, \dots, t_{1,K}\}$

t_1



Asymptotics for HDLSS Data

$\mathbf{x} = (x_1, \dots, x_d)^\top$ and $\mathbf{x}^o = (x_1^o, \dots, x_d^o)^\top$ represent non-outliers and outliers, respectively.

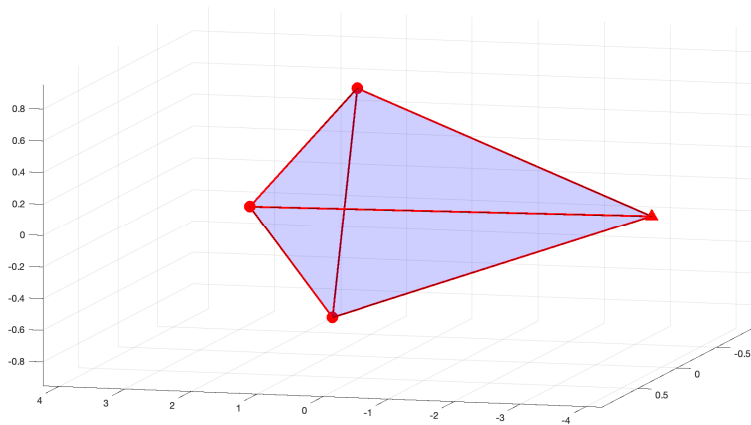
Assumptions

- (a) The fourth moments of the entries of the data vectors are uniformly bounded.
- (b) $d^{-1} \sum^d \{E(x_l) - E(x_l^o)\}^2 \rightarrow \mu^2$ as $d \rightarrow \infty$
- (c) $d^{-1} \sum^d \text{Var}(x_l) \rightarrow \sigma^2$ as $d \rightarrow \infty$
- (d) $d^{-1} \sum^d \text{Var}(x_l^o) \rightarrow \tau^2$ as $d \rightarrow \infty$
- (e) For both \mathbf{x} and \mathbf{x}^o , there exists a permutation of entries such that the sequence of the variables are ρ -mixing for functions that are dominated by quadratics.

The ρ -mixing in (e) is a mild condition to achieve the law of large numbers for sequence of correlated random variables.

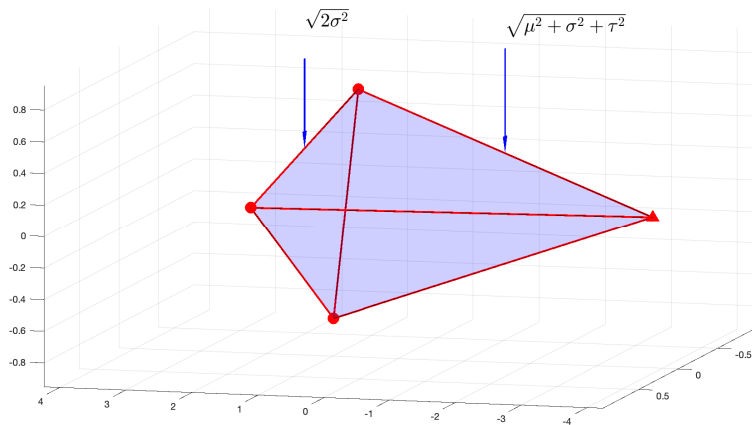
HDLSS geometric representation

- $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \stackrel{iid}{\sim} N(0_d, I_d)$ and $\mathbf{x}_4 \sim N(41_d, I_d)$, where $d = 5000$.



HDLSS geometric representation

- $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \stackrel{iid}{\sim} N(0_d, I_d)$ and $\mathbf{x}_4 \sim N(41_d, I_d)$, where $d = 5000$.



Asymptotic Sure Screening Probability

Theorem 4

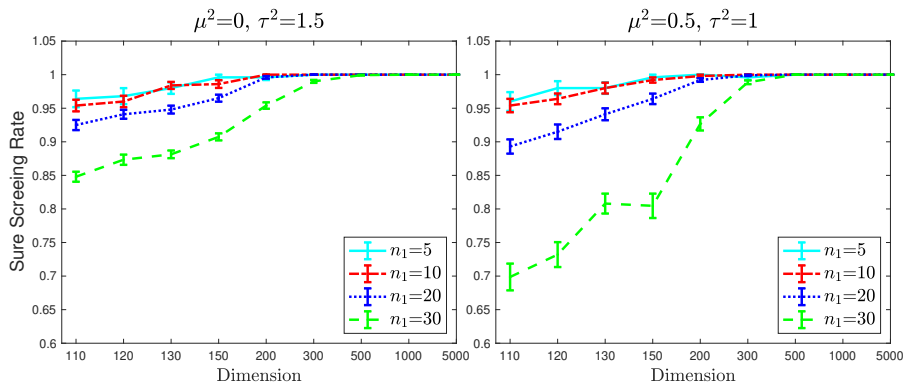
Let \mathcal{J} be the set of true outliers. Under the conditions, we have

$$\lim_{d \rightarrow \infty} \Pr(\mathcal{J} \subset \mathcal{X}_1) = 1,$$

if $\mu^2 + \tau^2 > \sigma^2$ and $n_0 > n_1 + 1$.

Sure Screening of Non-outliers

($N = 100$)



Asymptotic Power of SR Tests for Outlier Detection

Theorem 5

Suppose that $\mu^2 + \tau^2 > \sigma^2$, and let $\vartheta_t(\mathbf{Y})$ is the p-value such that

$$\vartheta_t(\mathbf{Y}) = E_{\mathbf{L}_n}[1\{t(\mathbf{Y}) \leq t(\mathbf{L}_n\mathbf{Y})\}|\mathbf{Y}].$$

Then, as d tends to infinity,

$$\vartheta_t(\mathbf{Y}) \xrightarrow{p} \begin{cases} 1 & \text{under } H_0 \\ 0 & \text{under } H_a \end{cases}$$

Asymptotic Power of SR Tests for Outlier Detection

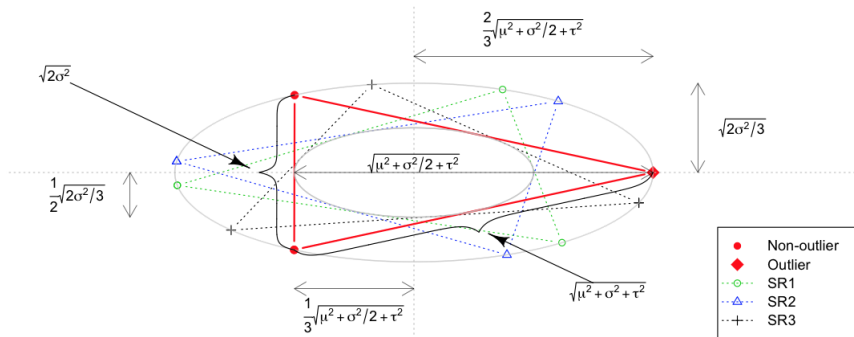


Illustration of the asymptotic geometry of HDLSS data and three randomly rotated data sets when $n_0 = 2$ and $n_1 = 1$.

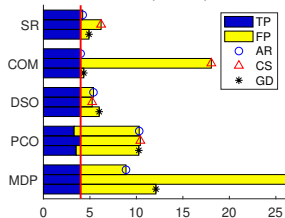
Simulation Study

- Normal, t , Gaussian Copula with exponential marginal
- $d = 1000$, $N = 50, 100$ and $n_1 = 0, 4, 15$.
- $\kappa = 500$, $\eta = 0.25$ and $\alpha = 0.05$.
- We consider three choices of covariance matrices as follows.
 - Auto-Regressive (AR): $\Sigma = \{0.8^{|l-l'|}\}_{l,l'}$, where $1 \leq l, l' \leq d$.
 - Compound Symmetry (CS): $\Sigma = .7I_d + .3J_{d,d}$.
 - Geometric Decaying (GD): $\Sigma = \Gamma\Lambda\Gamma^T$, where Γ is generated from $\text{Unif}(\mathcal{V}_{d,d})$ and Λ is a diagonal matrix with geometrically decaying eigenvalues. Specifically,

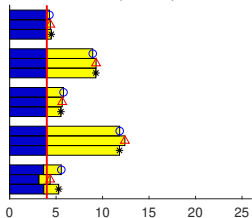
$$\lambda_l = \frac{d(.9^{l-1} - .9^l)}{1 - .9^d}, \quad l = 1, \dots, d.$$

Simulation Study

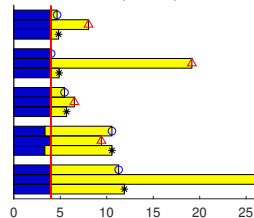
Location, N=100, MN



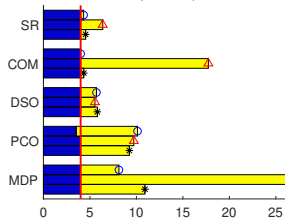
Location, N=100, MT



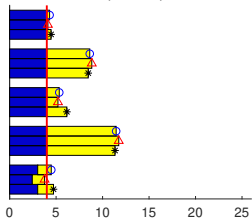
Location, N=100, GM



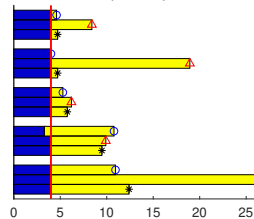
Scale, N=100, MN



Scale, N=100, MT

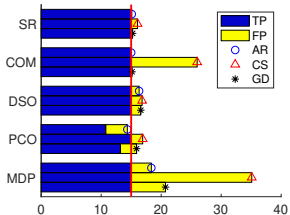


Scale, N=100, GM

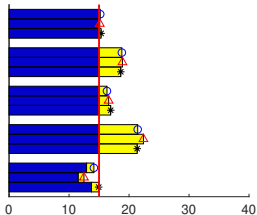


Simulation Study

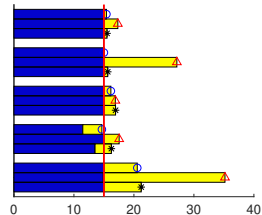
Location, N=100, MN



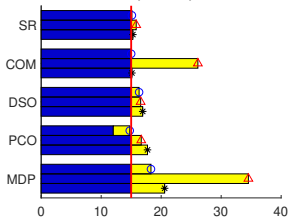
Location, N=100, MT



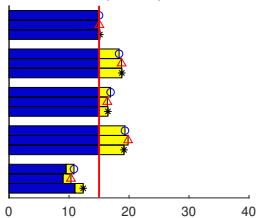
Location, N=100, GM



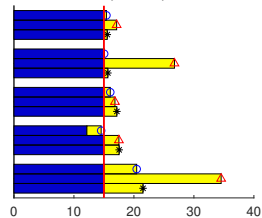
Scale, N=100, MN



Scale, N=100, MT



Scale, N=100, GM



Simulation Study (No outlier)

	N=50			N=100			
		AR	CS	GD	AR	CS	GD
MN	SR	0.44 (0.10)	2.48 (0.27)	0.50 (0.10)	0.54 (0.12)	2.74 (0.30)	0.84 (0.15)
	DSO	1.48 (0.15)	1.72 (0.25)	1.66 (0.14)	1.62 (0.16)	1.70 (0.19)	2.40 (0.31)
	PCO	4.84 (0.42)	4.66 (0.35)	4.80 (0.37)	7.94 (0.53)	7.40 (0.47)	8.14 (0.47)
	MDP	3.30 (0.25)	18.96 (0.43)	5.40 (0.39)	5.76 (0.39)	34.06 (0.78)	7.78 (0.42)
	COM	1.00 (0.00)	7.62 (0.38)	1.26 (0.07)	1.00 (0.00)	14.52 (0.49)	1.12 (0.05)
MT	SR	0.54 (0.11)	0.44 (0.11)	0.54 (0.12)	0.44 (0.08)	0.52 (0.12)	0.44 (0.11)
	DSO	1.70 (0.21)	1.64 (0.17)	1.60 (0.14)	2.20 (0.24)	2.36 (0.29)	2.58 (0.34)
	PCO	5.36 (0.35)	5.90 (0.42)	5.70 (0.41)	10.58 (0.61)	9.70 (0.57)	10.02 (0.50)
	MDP	1.68 (0.28)	1.32 (0.22)	1.72 (0.28)	2.18 (0.27)	1.68 (0.24)	2.08 (0.30)
	COM	6.06 (0.31)	6.80 (0.39)	6.40 (0.33)	5.56 (0.37)	5.98 (0.32)	5.54 (0.31)
GM	SR	0.68 (0.12)	3.36 (0.30)	0.78 (0.13)	0.88 (0.14)	5.22 (0.38)	1.30 (0.17)
	DSO	1.40 (0.16)	1.96 (0.22)	1.88 (0.22)	1.88 (0.23)	2.70 (0.41)	2.14 (0.22)
	PCO	4.98 (0.38)	5.12 (0.38)	5.62 (0.39)	7.34 (0.42)	7.94 (0.44)	8.06 (0.46)
	MDP	5.26 (0.39)	20.30 (0.33)	6.64 (0.48)	8.78 (0.47)	36.04 (0.86)	9.98 (0.54)
	COM	1.00 (0.00)	8.58 (0.33)	1.60 (0.11)	1.00 (0.00)	16.62 (0.47)	1.60 (0.12)

Outliers in Human Face Image Data

ORL face image data

- The ORL data consist of 400 images: 10 images of 40 individuals.
- An image consists of 112×92 pixels, $d = 10304$, with 0-255 grey levels per pixel.
- Created data with 10 images of one person plus 3 images of different people ($N = 13$).

	SR	DSO	PCO	COM
TPR	0.917	0.308	0.700	0.850
FPR	0.033	0.008	0.153	0.138

Outliers in Human Face Image Data



Conclusions

In this work,

- We developed an effective high dimensional outlier detection method.
- The proposed method controls type I error rate with the finite dimension and sample size.
- Also, the power of the proposed testing procedure converges to 1 as the dimension increases.
- Simulated and real data examples support our theoretical findings.

Future Directions

- Further look at the distribution assumption.
- Abnormality detection in different domains such as functional data and non-Euclidean data.
- Application of rotation test to other statistical problems.

Thank you for your attention!