

DOD: Detection of outliers in high dimensional data with distance of distances

Seong-ho Lee¹, Yongho Jeon^{2,3*}

¹Department of Statistics, University of Seoul, South Korea

²Department of Statistics and Data Science, Yonsei University, South Korea

³Department of Applied Statistics, Yonsei University, South Korea

Abstract

Reliable outlier detection in high-dimensional data is crucial in modern science, yet it remains a challenging task. Traditional methods often break down in these settings due to their reliance on asymptotic behaviors with respect to sample size under fixed dimension. Furthermore, many modern alternatives introduce sophisticated statistical treatments and computational complexities. To overcome these issues, our approach leverages intuitive geometric properties of high-dimensional space, effectively turning the curse of dimensionality into an advantage. We propose two new outlyingness statistics based on observation's relational patterns with all other points, measured via pairwise distances or inner products. We establish a theoretical foundation for our statistics demonstrating that as the dimension grows, our statistics create a non-vanishing margin that asymptotically separates outliers from non-outliers. Based on this foundation, we develop practical outlier detection procedures, including a simple clustering-based algorithm and a distribution-free test using random rotations. Through simulation experiments and real data applications, we demonstrate that our proposed methods achieve a superior balance between detection power and false positive control, outperforming existing methods and establishing their practical utility in high-dimensional settings.

Keywords: Outlier detection; high dimensional data; high dimensional asymptotics; data perturbation; random rotation

*corresponding author; email: yhjeon@yonsei.ac.kr

1 Introduction

The proliferation of high-dimensional data across scientific and industrial domains, from genomics and medical imaging to financial markets, has established outlier detection as a crucial analytical task. In these fields, outliers are often not merely noise to be discarded, but can represent the primary objects of interest, such as rare genetic variants associated with a disease, fraudulent financial transactions, or critical system failures (Chandola et al. 2009). This task is particularly crucial in high-dimension, low-sample-size (HDLSS) settings, where the number of features p vastly exceeds the number of observations n . In such scenarios, even a single outlier can cause serious distortions in statistical analysis, underscoring the need for robust and effective outlier detection methods.

Traditionally, a wide array of outlier detection methods were developed under the classical low-dimension, high-sample-size paradigm. These include methods based on distributional assumptions or approximations (McGill et al. 1978, Ye & Chen 2001), density-based clustering (Ester et al. 1996), and nearest-neighbor heuristics (Breunig et al. 2000). Many of these classical methods rely on metrics, such as the Mahalanobis distance, that summarize the data’s multivariate distribution to detect observations that deviate from the norm. However, the performance of these metrics degrades severely in high dimensions due to the “curse of dimensionality” (Zimek et al. 2012). The core of the problem lies in their reliance on the large sample asymptotics under fixed dimension. For instance, for the estimation of the covariance matrix, especially in HDLSS settings where $p \gg n$, the sample covariance matrix is singular and cannot be inverted, or its estimate is subject to high variability. This statistical and numerical instability causes the collapse of classical metrics, rendering many well-established outlier detection methods ineffective or entirely inapplicable for high-dimensional data.

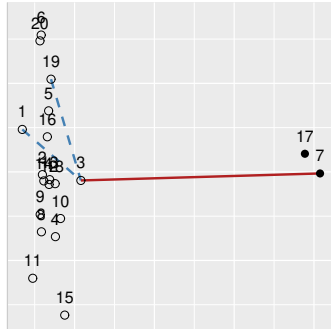
In response to these challenges, a new generation of methods designed specifically for high-dimensional data has emerged. These methods include approaches based on measuring local density variation (Papadimitriou et al. 2003), using angles instead of distances

(Kriegel et al. 2008), and developing robust versions of classical methods. For instance, Filzmoser et al. (2008) proposed a reweighting method using principal components, while Ro et al. (2015) improved the Mahalanobis distance by using the minimum covariance determinant. More recently, Chung & Ahn (2021) introduced a metric based on a distance to hyperplane, and devised a two-stage procedure that conducts a hypothesis test for each outlier candidate. A key advantage of these methods is that they can operate without relying on the large sample asymptotics. However, this advantage comes with its own cost; many of these modern alternative methods require sophisticated statistical treatments thus computationally complex, along with careful tuning parameter settings.

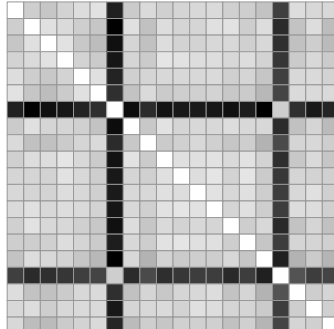
To overcome these issues, we propose a new outlier detection method that is computationally simple while theoretically grounded. Our approach draws inspiration from the concept of distance vector clustering introduced by Terada (2013), which was shown to be an efficient alternative to methods based on the maximal data piling direction (Ahn & Marron 2010, Ahn et al. 2012), as it discriminates groups based on metrics that are simple yet effectively reflect both mean and variance differences in different groups. Instead of relying on complex statistical treatments, we innovate this simple concept to devise an outlyingness statistic for individual observations by leveraging intuitive geometric properties of high-dimensional space. The core insight is that outliers exhibit a relational pattern with respect to all other data points that is fundamentally different from that of non-outliers (or inliers). We capture this characteristic by adopting the concepts of *Distance of Distances* (DOD) and *Distance of inner products in a Gram matrix* (DOG). These concepts lead to two new outlyingness measure statistics which quantify how much an observation’s entire profile of pairwise relationships deviates from the typical profile of non-outlying points.

The contributions of this paper are three-fold. First, we establish a theoretical foundation for our statistics, demonstrating that as the dimension p grows, the statistics create a non-vanishing asymptotic margin between outliers and non-outliers. Second, based on this theoretical foundation, we develop a set of practical outlier detection procedures, including

Illustration of the proposed outlier detection statistic. Panel (a) shows a 2D of a simulated dataset containing two outliers (7 and 17). Panel (b) shows the Δ_D . Panel (c) shows the barplot of $t_i^{(D)}$, which is markedly larger for the outliers.



(a) Data scatterplot



(b) Heatmap of Δ_D

(c) Barplot of $t_i^{(D)}$

a simple clustering-based algorithm, and data-driven non-parametric tests based on random rotations (Blaser & Fryzlewicz 2016), a technique effectively used in Chung & Ahn (2021). Third, we demonstrate through simulation experiments and applications to two real datasets, a microarray gene expression dataset and a human face image dataset, that our methods achieve a superior balance between high detection power and stringent false positive control compared to existing methods.

The rest of the paper is organized as follows. Section 2 introduces the proposed statistics and their theoretical properties. Section 3 details the outlier detection procedures. Sections 4 and 5 present the numerical results from simulation experiments and real data applications, respectively. Finally, Section 6 concludes the paper.

2 Proposed statistics and theoretical properties

2.1 Proposed statistics

A core motivation for our proposed statistics stems from the observation of how outliers manifest in pairwise relationship matrices. Figure 1 illustrates this phenomenon based on a simulated dataset with dimension $p = 1000$ and sample size $n = 20$, containing two dis-

nated outliers (the 7th and 17th observations). In the 2D projection shown in Figure 1(a), a key visual takeaway is that the length of the solid line, representing the dissimilarity between an inlier and an outlier, is substantially greater than that of the dashed line, representing the dissimilarity between two inliers.

By aggregating this relational information for all pairs, we construct an $n \times n$ matrix Δ_D visualized as a heatmap in Figure 1(b), which will be elaborated in a sequel. Large dissimilarities between observations appear as dark entries in the heatmap. These dark entries form distinct, high-magnitude columns that correspond to the outliers. This clear pattern demonstrates that an outlier’s relational profile is profoundly different from that of a non-outlier, which leads to a critical insight: the column-wise median of this matrix can serve as a robust baseline for the typical relational pattern, and consequently, a substantial deviation from this baseline can serve as a strong indicator of outlyingness.

This phenomenon directly underpins the design of our proposed statistics $t_i^{(D)}$ and $t_i^{(G)}$. We now elaborate our first proposed statistic $t_i^{(D)}$. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a centered data matrix with its i th row denoted as $\mathbf{x}_i \in \mathbb{R}^p$. We start by computing the distance matrix $\mathbf{D} \in \mathbb{R}^{n \times n}$, where $[\mathbf{D}]_{i,j} = d(\mathbf{x}_i, \mathbf{x}_j)$. For ease of presentation, we use the Euclidean distance for $d(\cdot, \cdot)$. Based on this matrix, we construct the matrix of distances-of-distances, $\Delta_D \in \mathbb{R}^{n \times n}$ with elements $[\Delta_D]_{i,j} = \delta_{ij}^{(D)} = \sqrt{\sum_{k \neq i,j} ([\mathbf{D}]_{i,k} - [\mathbf{D}]_{j,k})^2}$. $\delta_{ij}^{(D)}$ measures the dissimilarity between the distance patterns of $\mathbf{x}_i$ and $\mathbf{x}_j$ relative to all other observations. Following this construction, we define our first statistic $t_i^{(D)}$ as the Euclidean distance between the i th row of Δ_D and the column-wise median vector of Δ_D . The column-wise median vector serves as a robust representation of the typical pattern of non-outlying points. Formally,

$$t_i^{(D)} = \sqrt{\sum_{j=1}^n \left\{ \delta_{ij}^{(D)} - \tilde{\delta}_{\cdot j}^{(D)} \right\}^2},$$

where $\tilde{\delta}_{\cdot j}^{(D)} = \text{median}\{\delta_{ij}^{(D)} : i = 1, \dots, n\}$.

Similarly, we propose a second statistic $t_i^{(G)}$ based on inner products, which captures

different aspects of dissimilarity from $t_i^{(D)}$. We first compute the inner product matrix $\mathbf{G} \in \mathbb{R}^{n \times n}$, with elements $[\mathbf{G}]_{i,j} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$. We focus on the dot product. Analogous to Δ_D , we construct the matrix of distances-of-inner-products $\Delta_G \in \mathbb{R}^{n \times n}$ with elements $[\Delta_G]_{i,j} = \delta_{ij}^{(G)} = \sqrt{\sum_{k \neq i,j} ([\mathbf{G}]_{i,k} - [\mathbf{G}]_{j,k})^2}$. This quantity $\delta_{ij}^{(G)}$ measures the dissimilarity in inner product patterns between $\mathbf{x}_i$ and $\mathbf{x}_j$. Our second statistic $t_i^{(G)}$ is defined as the Euclidean distance between the i th row of Δ_G and its column-wise median vector:

$$t_i^{(G)} = \sqrt{\sum_{j=1}^n \left\{ \delta_{ij}^{(G)} - \tilde{\delta}_{\cdot j}^{(G)} \right\}^2},$$

where $\tilde{\delta}_{\cdot j}^{(G)} = \text{median}\{\delta_{ij}^{(G)} : i = 1, \dots, n\}$.

To illustrate the consequence of this design, Figure 1(c) displays a barplot of the proposed statistic $t_i^{(D)}$ computed under the same simulated dataset for Figures 1(a) and 1(b). Consistent with the patterns of the heatmap in Figure 1(b), the statistics corresponding to the two outliers—the 7th and 17th observations—are markedly larger than those for the remaining non-outlying points. This stark separation of magnitudes provides an empirical demonstration of the utility of our proposed statistics in outlier detection. The theoretical justification for this phenomenon, which guarantees a clear margin between the outlier and non-outlier statistics in high-dimensional settings, is established in Theorems 1 and 2 in the following section.

2.2 Theoretical properties

To rigorously validate the effectiveness of our proposed statistics $t_i^{(D)}$ and $t_i^{(G)}$ under high-dimensional settings, we now establish the theoretical properties of the proposed statistics under a set of assumptions. Let $\mathbf{x}^{(I)} = [X_1^{(I)}, \dots, X_p^{(I)}]^\top \in \mathbb{R}^p$ be the random vector representing non-outliers and $\mathbf{x}^{(O)} = [X_1^{(O)}, \dots, X_p^{(O)}]^\top \in \mathbb{R}^p$ be the random vector representing outliers. Following the framework of Hall et al. (2005), a common approach in high-dimensional asymptotic studies, we assume the following conditions.

(H1) The fourth moments of the entries of the sample vectors are uniformly bounded.

(H2) $\lim_{p \rightarrow \infty} \frac{1}{p} \sum_{k=1}^p \mathbb{E}\{X_k^{(I)}\}^2 = \mu_I^2$ and $\lim_{p \rightarrow \infty} \frac{1}{p} \sum_{k=1}^p \mathbb{E}\{X_k^{(O)}\}^2 = \mu_O^2$.

(H3) $\lim_{p \rightarrow \infty} \frac{1}{p} \sum_{k=1}^p \mathbb{V}\{X_k^{(I)}\} = \sigma_I^2$ and $\lim_{p \rightarrow \infty} \frac{1}{p} \sum_{k=1}^p \mathbb{V}\{X_k^{(O)}\} = \sigma_O^2$.

(H4) $\lim_{p \rightarrow \infty} \frac{1}{p} \sum_{k=1}^p \left[\mathbb{E}\{X_k^{(I)}\} - \mathbb{E}\{X_k^{(O)}\} \right]^2 = \delta^2$.

(H5) For all random vectors, there exists a permutation of entries such that the sequence of the variables are ρ -mixing for functions that are dominated by quadratics.

These conditions provide a foundation to analyze the limiting behavior of the proposed statistics. Specifically, Conditions (H2) and (H3) ensure that the per-feature mean squared expectation and variance of both non-outliers and outliers converge to fixed values as the dimension p grows. This allows for a stable characterization of each group. Condition (H4) is also crucial in that it formalizes the separation between the non-outlier and outlier clusters, ensuring that the squared mean difference between the two groups does not vanish in the high-dimensional limit.

To further provide a theoretical support for the proposed statistics, we state the asymptotic behavior of the constituent quantities of Δ_D and Δ_G as $p \rightarrow \infty$. The following lemma, a corrected and restated version of the result from Terada (2013) under the assumptions from Hall et al. (2005), provides the asymptotic limits for the pairwise differences in distances and inner products.

Lemma 1 *Under Conditions (H1)–(H5), we have the following results as $p \rightarrow \infty$:*

(i) *If either $\mathbf{x}_i, \mathbf{x}_j \sim \mathbf{x}^{(I)}$ or $\mathbf{x}_i, \mathbf{x}_j \sim \mathbf{x}^{(O)}$,*

$$\begin{aligned} \frac{1}{\sqrt{p}}([\mathbf{D}]_{i,k} - [\mathbf{D}]_{j,k}) &\xrightarrow{p} 0, \\ \frac{1}{p}([\mathbf{G}]_{i,k} - [\mathbf{G}]_{j,k}) &\xrightarrow{p} 0. \end{aligned}$$

(ii) If $\mathbf{x}_i \sim \mathbf{x}^{(I)}$ and $\mathbf{x}_j \sim \mathbf{x}^{(O)}$,

$$\begin{aligned} \frac{1}{\sqrt{p}}([\mathbf{D}]_{i,k} - [\mathbf{D}]_{j,k}) &\xrightarrow{p} \begin{cases} \sqrt{2}\sigma_I - \sqrt{\sigma_I^2 + \sigma_O^2 + \delta^2} := \alpha_D & \text{if } \mathbf{x}_k \sim \mathbf{x}^{(I)}, \\ \sqrt{\sigma_I^2 + \sigma_O^2 + \delta^2} - \sqrt{2}\sigma_O := \beta_D & \text{if } \mathbf{x}_k \sim \mathbf{x}^{(O)}. \end{cases} \\ \frac{1}{p}([\mathbf{G}]_{i,k} - [\mathbf{G}]_{j,k}) &\xrightarrow{p} \begin{cases} \frac{\mu_I^2 - \mu_O^2 + \delta^2}{2} := \alpha_G & \text{if } \mathbf{x}_k \sim \mathbf{x}^{(I)}, \\ \frac{\mu_I^2 - \mu_O^2 - \delta^2}{2} := \beta_G & \text{if } \mathbf{x}_k \sim \mathbf{x}^{(O)}. \end{cases} \end{aligned}$$

(iii) We have $\alpha_D = \beta_D = 0$ if and only if $\sigma_I^2 = \sigma_O^2$ and $\delta^2 = 0$. Also, $\alpha_G = \beta_G = 0$ if and only if $\mu_I^2 = \mu_O^2$ and $\delta^2 = 0$.

Lemma 1 suggests that these pairwise differences converge to distinct, non-zero values depending on whether the observations involved are non-outliers or outliers, which is a critical property for our statistics to effectively differentiate the two groups. Specifically, Lemma 1(iii) underscores the distinct characteristics of the two statistics; the distance-based measure is primarily sensitive to the discrepancy in population variances σ_I^2 and σ_O^2 , while the inner-product-based measure captures the difference in squared norms μ_I^2 and μ_O^2 .

Based on the asymptotic behaviors established in Lemma 1, we now present the main theoretical result concerning our proposed statistics. The following theorem formally demonstrates that our statistics can effectively distinguish between non-outliers and outliers in the high-dimensional setting. Its proof is provided in the supplementary material.

Theorem 1 *Under Conditions (H1)–(H5), we have the following results as $p \rightarrow \infty$:*

(i) *For a non-outlier $\mathbf{x}_i \sim \mathbf{x}^{(I)}$, the scaled statistics converge to zero in probability:*

$$\frac{1}{\sqrt{p}}t_i^{(D)} \xrightarrow{p} 0 \quad \text{and} \quad \frac{1}{p}t_i^{(G)} \xrightarrow{p} 0.$$

(ii) For an outlier $\mathbf{x}_i \sim \mathbf{x}^{(O)}$, the scaled statistics converge to constants in probability:

$$\begin{aligned} \frac{1}{\sqrt{pn}} t_i^{(D)} &\xrightarrow{p} \sqrt{(n - n_{\text{out}} - 1)\alpha_D^2 + (n_{\text{out}} - 1)\beta_D^2} &:= \gamma_D, \\ \frac{1}{p\sqrt{n}} t_i^{(G)} &\xrightarrow{p} \sqrt{(n - n_{\text{out}} - 1)\alpha_G^2 + (n_{\text{out}} - 1)\beta_G^2} &:= \gamma_G. \end{aligned}$$

Remark 1 (Individual Distinction) *Theorem 1 demonstrates a distinction in the asymptotic behavior of the proposed statistics for non-outliers versus outliers. For any non-outlying observation, the scaled statistic is asymptotically negligible, as its value vanishes toward zero in the high-dimensional limit. In contrast, the scaled statistic for an outlier captures a signal of its anomalous nature, converging to a positive constant. This divergent limiting behavior provides a theoretical support for their distinction by adopting our proposed statistics.*

Remark 2 (Enhanced Detection from Sample Size n and Dimensionality p) *The performance of our proposed statistics is enhanced by both sample size n and dimensionality p . Firstly, a larger sample size n directly magnifies the statistic $t_i^{(D)}$ for outliers, while leaving it unchanged for non-outliers. Specifically, for an outlier, the magnitude of $t_i^{(D)}$ grows linearly with n since $t_i^{(D)} \approx \sqrt{pn} \gamma_D$ and $\gamma_D \propto \sqrt{n}$. In contrast, for a non-outlier, its magnitude $o_p(\sqrt{p})$ is independent of n . This creates a widening gap between the outlier and non-outlier statistics as the sample size increases, thereby strengthening detection power.*

Secondly, our method leverages high dimensionality. It relies on the convergent behavior of pairwise distances and inner products in high-dimensional spaces, where large p ensures the asymptotic stabilization described in the conditions and theorems. This dimension leveraging effectively turns the classic curse of dimensionality into an advantage for outlier detection, making our proposed statistics useful for high-dimensional data.

Building upon the individual convergence properties shown in Theorem 1, we now advance to a stronger, collective statement. While the previous theorem guarantees that the individual statistic for any non-outlier vanishes while that for an outlier remains large, it yet

does not preclude the possibility of overlap between the two populations. The next theorem resolves this issue by proving that a non-vanishing margin indeed exists, separating the entire set of outliers from the set of non-outliers. Its proof is provided in the supplementary material.

Theorem 2 *Let $\mathcal{I}$ and $\mathcal{O}$ be the index sets for non-outliers and outliers, respectively. Under Conditions (H1)–(H5), the gap between the scaled outlier and non-outlier statistics converges to constants in probability as $p \rightarrow \infty$:*

$$\begin{aligned} \min_{i \in \mathcal{O}} \frac{t_i^{(D)}}{\sqrt{pn}} - \max_{i \in \mathcal{I}} \frac{t_i^{(D)}}{\sqrt{pn}} &\xrightarrow{p} \gamma_D, \\ \min_{i \in \mathcal{O}} \frac{t_i^{(G)}}{p\sqrt{n}} - \max_{i \in \mathcal{I}} \frac{t_i^{(G)}}{p\sqrt{n}} &\xrightarrow{p} \gamma_G. \end{aligned}$$

Corollary 1 *Under Conditions (H1)–(H5), the gap between the scaled outlier and non-outlier statistics is bounded strictly above zero in probability as $p \rightarrow \infty$:*

$$\begin{aligned} \lim_{p \rightarrow \infty} \Pr \left\{ \min_{i \in \mathcal{O}} \frac{t_i^{(D)}}{\sqrt{pn}} - \max_{i \in \mathcal{I}} \frac{t_i^{(D)}}{\sqrt{pn}} > 0 \right\} &= 1 \quad \text{if } \sigma_I^2 \neq \sigma_O^2 \text{ or } \delta^2 \neq 0, \\ \lim_{p \rightarrow \infty} \Pr \left\{ \min_{i \in \mathcal{O}} \frac{t_i^{(G)}}{p\sqrt{n}} - \max_{i \in \mathcal{I}} \frac{t_i^{(G)}}{p\sqrt{n}} > 0 \right\} &= 1 \quad \text{if } \mu_I^2 \neq \mu_O^2 \text{ or } \delta^2 \neq 0. \end{aligned}$$

Remark 3 (Existence of a Separation Margin) *Theorem 2 and Corollary 1 provide a stronger theoretical guarantee for our statistics' outlier detection performance. It demonstrates that as the dimension grows, the two groups become perfectly separated; the smallest scaled statistic from the outlier group becomes strictly greater than the largest scaled statistic from the non-outlier group. This result offers a justification for distinguishing outliers, as a clear margin emerges between the two populations. The existence of this non-vanishing separation margin γ_D (or γ_G) ensures that the detection capability of the proposed statistics is reliable in high-dimensional settings.*

3 Proposed outlier detection procedure

3.1 Detection via clustering

Our theoretical results provide a foundation for a practical detection procedure. The key insight stems from Theorem 2, which guarantees that as the dimension p grows, a non-vanishing margin emerges between the scaled statistics of non-outliers and outliers. This asymptotic separability is the cornerstone of our proposed procedure, as it effectively transforms the complex, high-dimensional outlier detection problem into a much simpler, one-dimensional clustering task performed on the set of statistics $\{t_i\}_{i=1}^n$.

Leveraging this theoretical guarantee, we propose a straightforward outlier detection procedure via clustering. The procedure begins by computing outlyingness statistics, either $t_i^{(D)}$ or $t_i^{(G)}$, for each observation $\mathbf{x}_i$ in the dataset. Subsequently, a standard clustering algorithm is applied to partition these n statistics into two distinct groups. Given the clear separation shown by Theorem 2, a simple algorithm such as k-means is sufficient to effectively distinguish the two populations.

The final step is to label the two clusters and validate their separation. The cluster with the larger mean statistic is designated as the potential outlier group, C_{out} . To avoid the pitfall of wrongly declaring this cluster as outliers in outlier-free scenarios, we validate the separation between the groups. We compute the gap defined as $g = \min_{i \in C_{\text{out}}} t_i - \max_{j \in C_{\text{in}}} t_j$, where C_{in} is the non-outlier cluster. The members of C_{out} are then declared as outliers only if this gap exceeds a predefined gap threshold $c > 0$ and if the cluster's size $|C_{\text{out}}|$ is less than a specified proportion of the total sample size $n\alpha$. The parameter $\alpha \in (0, 0.5)$ represents the maximum proportion of outliers, thus serves as a tuning parameter controlling the maximum false positive rate (FPR). Otherwise, we conclude that no distinct group of outliers exists and return an empty set. This procedure is summarized in Algorithm 1.

Algorithm 1 Outlier Detection via Clustering

Input: Centered data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, maximum FPR α , gap threshold c .

Output: Index set of declared outliers $\hat{\mathcal{O}}$.

- 1: Compute the statistics $\{t_1, \dots, t_n\}$ from $\mathbf{X}$.
 - 2: Partition $\{t_1, \dots, t_n\}$ into two clusters C_1 and C_2 using a clustering algorithm.
 - 3: Identify the potential outlier cluster C_{out} as the cluster with the larger mean of the statistic, and C_{in} as the other.
 - 4: Compute the gap: $g = \min_{i \in C_{\text{out}}} t_i - \max_{j \in C_{\text{in}}} t_j$.
 - 5: **if** $|C_{\text{out}}| \leq n\alpha$ and $g > c$ **then**
 - 6: $\hat{\mathcal{O}} \leftarrow \{i \mid i \in C_{\text{out}}\}$.
 - 7: **else**
 - 8: $\hat{\mathcal{O}} \leftarrow \emptyset$.
 - 9: **end if**
 - 10: **Return** $\hat{\mathcal{O}}$.
-

3.2 Detection via random rotation

As an alternative to clustering for outlier detection, we propose a non-parametric testing procedure based on random rotation (Blaser & Fryzlewicz 2016). Random rotation is a data perturbation technique where, for a data $\mathbf{X}$, a rotated version $\mathbf{X}^* = \mathbf{H}\mathbf{X}$ is generated by pre-multiplying a randomly sampled rotation matrix $\mathbf{H}$. This allows us to generate a reference or “null” distribution for a test statistic directly from the observed data, providing a distribution-free, data-driven decision boundary for hypothesis testing. Thus, instead of relying on a predefined gap threshold c as Algorithm 1, a new proposed procedure will provide a data-driven threshold.

The theoretical justification for our proposal is grounded in the properties of the left-spherical distribution family (Chung & Ahn 2021). Let us establish a null hypothesis H_0 that the non-outlier data follows a left-spherical distribution. Under this hypothesis, the non-outlier data distribution is invariant to pre-multiplication by any orthogonal matrix $\mathbf{H}$. Rotating the entire dataset by pre-multiplying $\mathbf{H}$ to $\mathbf{X}$, we can simulate the distribution of test statistics under the null hypothesis, as the rotation randomizes observation-specific quantities while preserving the distribution of the entire dataset. It is worthwhile to note that our proposed statistics $t_i^{(\text{D})}$ and $t_i^{(\text{G})}$ are dependent on the relative arrangement of the

Algorithm 2 Outlier Detection via Random Rotation

Input: Centered data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, maximum FPR α , number of rotations B .

Output: Index set of declared outliers $\hat{\mathcal{O}}$.

- 1: Compute the statistics $\{t_1, \dots, t_n\}$ from $\mathbf{X}$.
 - 2: Initialize an empty set for the null distribution: $\mathcal{T}_{\text{null}} \leftarrow \emptyset$.
 - 3: **for** $b = 1$ to B **do**
 - 4: Generate a random orthogonal matrix $\mathbf{H}_b$.
 - 5: Compute the rotated data matrix $\mathbf{X}_b = \mathbf{H}_b \mathbf{X}$.
 - 6: Compute the statistics $\{t_{1,b}, \dots, t_{n,b}\}$ from $\mathbf{X}_b$.
 - 7: Update the null distribution: $\mathcal{T}_{\text{null}} \leftarrow \mathcal{T}_{\text{null}} \cup \{t_{1,b}, \dots, t_{n,b}\}$.
 - 8: **end for**
 - 9: Determine the critical value c_α as the $(1 - \alpha)$ th quantile of $\mathcal{T}_{\text{null}}$.
 - 10: Identify the outlier index set $\hat{\mathcal{O}} = \{i \mid t_i > c_\alpha\}$.
 - 11: **Return** $\hat{\mathcal{O}}$.
-

observations therefore rotation-variant, making them well-suited for this procedure.

Specifically, the random rotation test for outlier detection can be implemented in two different ways, with the second offering superior statistical properties. A first, straightforward implementation of the random rotation test involves creating a null distribution by pooling all statistics from the rotated data. The procedure begins by computing the statistics $\{t_1, \dots, t_n\}$ for the original data. Subsequently, a number of rotated datasets $\mathbf{X}_b$ ($b = 1, \dots, B$) are generated, and the statistics $\{t_{1,b}, \dots, t_{n,b}\}$ are computed for each. All $n \times B$ of these statistics are then aggregated into a single empirical null distribution $\mathcal{T}_{\text{null}}$. The critical value c_α is then determined by the $(1 - \alpha)$ th quantile of this distribution. Finally, we declare the i th observation as an outlier if $t_i > c_\alpha$. This procedure is detailed in Algorithm 2.

While intuitive, this method fails to account for the multiple comparisons problem inherent in testing n hypotheses simultaneously. Consequently, the probability of making at least one false discovery is not controlled at the nominal level α , potentially leading to an inflated number of false positives. To address this shortcoming, we further propose a procedure that controls the Family-Wise Error Rate (FWER). The FWER is the probability of making one or more false discoveries, thus controlling it provides a much stronger guarantee of statistical validity.

Algorithm 3 Outlier Detection via Random Rotation with FWER Control

Input: Centered data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, maximum family-wise FPR α , number of rotations B .

Output: Index set of declared outliers $\hat{\mathcal{O}}$.

- 1: Compute the statistics $\{t_1, \dots, t_n\}$ from $\mathbf{X}$.
 - 2: Initialize an empty set for the null distribution: $\mathcal{T}_{\text{null}} \leftarrow \emptyset$.
 - 3: **for** $b = 1$ to B **do**
 - 4: Generate a random orthogonal matrix $\mathbf{H}_b$.
 - 5: Compute the rotated data matrix $\mathbf{X}_b = \mathbf{H}_b \mathbf{X}$.
 - 6: Compute the statistics $\{t_{1,b}, \dots, t_{n,b}\}$ from $\mathbf{X}_b$.
 - 7: Find the maximum statistic: $t_{\max,b} = \max_i \{t_{i,b}\}$.
 - 8: Update the null distribution: $\mathcal{T}_{\text{null}} \leftarrow \mathcal{T}_{\text{null}} \cup \{t_{\max,b}\}$.
 - 9: **end for**
 - 10: Determine the critical value c_α as the $(1 - \alpha)$ -th quantile of $\mathcal{T}_{\text{null}}$.
 - 11: Identify the outlier index set $\hat{\mathcal{O}} = \{i \mid t_i > c_\alpha\}$.
 - 12: **Return** $\hat{\mathcal{O}}$.
-

This is achieved by constructing the null distribution of the maximum statistic from each rotated data $t_{\max,b} = \max_i \{t_{i,b}\}$. The collection of these maximums forms an empirical null distribution of the most extreme statistic under H_0 . The resulting critical value c_α is consequently more conservative. This FWER-controlled procedure is particularly powerful when paired with our proposed statistics $t_i^{(\text{D})}$ and $t_i^{(\text{G})}$. As established in Theorems 1 and 2, our statistics for true outliers diverge and form a clear margin from the statistics of non-outliers. Therefore, even though the critical value c_α constructed from the FWER procedure is more conservative, we can expect that the statistics of true outliers reliably exceed this threshold, thus ensure high detection power while maintaining stringent error control. This procedure is formally described in Algorithm 3.

4 Simulation experiment

We conduct a simulation study to evaluate the empirical performance of our proposed outlier detection procedures. Our proposed procedures, denoted as DOD1, DOD2, DOD3 and DOG1, DOG2, DOG3, are based on implementing two different statistics $t_i^{(\text{D})}$ for DOD

and $t_i^{(G)}$ for DOG, with Algorithms 1, 2, and 3, respectively. We benchmark their performance against three competing methods: Subspace Rotation-based outlier detection by Chung & Ahn (2021) (SRout), Minimum Diagonal Product by Ro et al. (2015) (RMDP), and Principal Component-based outlier detection by Filzmoser et al. (2008) (PCout). We implemented all competing methods with their default parameters as provided by the original authors.

We simulate a data matrix $\mathbf{X}$ of size $n \times p$ with $n = 30$ and $p = 500$ containing n_{out} outliers. The generation of the $(n - n_{\text{out}})$ non-outlier observations depends on the specified structure. For the Identity (ID) and Auto-Regressive (AR) structures, the non-outliers are drawn from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\text{in}})$, where $\mathbf{\Sigma}_{\text{in}} = \mathbf{I}_p$ for the ID structure, and $[\mathbf{\Sigma}_{\text{in}}]_{j,k} = 0.7^{|j-k|}$ for the AR structure. For the Moving Average (MA) structure, non-outliers are generated directly from the process $X_j = \frac{\sum_{l=1}^L \eta_l Z_{j+l-1}}{(\sum_{l=1}^L \eta_l^2)^{1/2}}$ for $j = 1, \dots, p$, where the Z_k are independent standard normal variables, the coefficients η_l are drawn from a uniform distribution $\mathcal{U}(0, 1)$, and $L = \lfloor \sqrt{p} \rfloor$. In contrast, each outlier is independently generated from $\mathcal{N}(p^{s_\mu} \mathbf{u} / \|\mathbf{u}\|_2, s_\sigma \mathbf{I}_p)$, with elements of $\mathbf{u}$ drawn independently from $\mathcal{U}(0, 1)$. The parameter s_μ controls the mean shift magnitude, while s_σ scales the outlier covariance. Our simulations include scenarios with no outliers ($n_{\text{out}} = 0$), as well as with $n_{\text{out}} = 3$ under varying outlier magnitudes determined by (s_μ, s_σ) pairs of $(0.5, 1.0)$, $(0.5, 0.5)$, and $(0.25, 0.25)$.

The tuning parameters for our proposed procedures are set as follows. For the clustering-based method detailed in Algorithm 1, we use k-means for clustering with $k = 2$ and set the maximum allowable proportion of outliers to $\alpha = 0.3$. The gap threshold c is chosen to align with the asymptotic behavior of the test statistics as stated in Corollary 1. Specifically, we set $c = 0.1\sqrt{pn}$ for DOD1 and $c = 0.1p\sqrt{n}$ for DOG1. For the random rotation methods detailed in Algorithms 2 and 3, we generate $B = 300$ randomly rotated datasets. Further, we use $\alpha = 0.05$ for DOD2 and DOG2, and $\alpha = 0.7$ for DOD3 and DOG3.

Table 1 summarizes the simulation results from 1000 replicates for scenarios with three outliers ($n_{\text{out}} = 3$). We assess the performance of each method using three metrics: the

Table 1: Summary of simulation experiment under $n_{\text{out}} = 3$.

(s_μ, s_σ)	Method	ID			AR			MA		
		TPR	FPR	FWFP	TPR	FPR	FWFP	TPR	FPR	FWFP
(0.5, 1.0)	DOD1	1.000	0.000	0.000	1.000	0.000	0.001	0.996	0.006	0.141
	DOD2	1.000	0.000	0.000	1.000	0.002	0.065	1.000	0.020	0.423
	DOD3	1.000	0.000	0.000	1.000	0.001	0.024	1.000	0.011	0.248
	DOG1	1.000	0.000	0.000	1.000	0.000	0.000	0.831	0.006	0.132
	DOG2	1.000	0.000	0.000	1.000	0.000	0.000	0.981	0.002	0.041
	DOG3	1.000	0.000	0.000	1.000	0.000	0.000	0.912	0.001	0.021
	SRout	1.000	0.007	0.176	1.000	0.007	0.167	1.000	0.007	0.164
	RMDP	1.000	0.159	0.989	1.000	0.126	0.948	1.000	0.117	0.934
	PCout	1.000	0.047	0.692	0.996	0.055	0.763	0.793	0.081	0.860
(0.5, 0.5)	DOD1	1.000	0.000	0.000	0.993	0.002	0.061	0.634	0.019	0.332
	DOD2	1.000	0.000	0.001	1.000	0.004	0.096	0.881	0.020	0.420
	DOD3	1.000	0.000	0.000	1.000	0.002	0.042	0.732	0.011	0.255
	DOG1	1.000	0.000	0.000	1.000	0.000	0.001	0.833	0.006	0.133
	DOG2	1.000	0.000	0.000	1.000	0.000	0.000	1.000	0.005	0.134
	DOG3	1.000	0.000	0.000	1.000	0.000	0.000	0.992	0.002	0.057
	SRout	1.000	0.007	0.176	0.941	0.007	0.169	0.480	0.007	0.166
	RMDP	1.000	0.161	0.985	1.000	0.127	0.950	0.965	0.118	0.926
	PCout	1.000	0.045	0.677	0.977	0.062	0.792	0.597	0.092	0.859
(0.25, 0.25)	DOD1	1.000	0.000	0.000	1.000	0.000	0.004	0.958	0.019	0.355
	DOD2	1.000	0.009	0.212	1.000	0.046	0.716	1.000	0.079	0.905
	DOD3	1.000	0.005	0.126	1.000	0.037	0.613	1.000	0.063	0.827
	DOG1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.033
	DOG2	0.000	0.000	0.000	0.000	0.022	0.464	0.033	0.064	0.922
	DOG3	0.000	0.000	0.000	0.000	0.014	0.323	0.014	0.045	0.792
	SRout	0.000	0.119	0.952	0.000	0.088	0.915	0.000	0.063	0.808
	RMDP	0.000	0.048	0.671	0.000	0.100	0.911	0.000	0.165	0.981
	PCout	0.990	0.055	0.745	0.957	0.050	0.702	0.746	0.071	0.814

Table 2: Summary of simulation experiment under $n_{\text{out}} = 0$.

Method	ID		AR		MA	
	FPR	FWFP	FPR	FWFP	FPR	FWFP
DOD1	0.001	0.027	0.005	0.108	0.016	0.246
DOD2	0.009	0.224	0.033	0.634	0.045	0.755
DOD3	0.004	0.105	0.026	0.538	0.035	0.655
DOG1	0.000	0.000	0.000	0.000	0.001	0.029
DOG2	0.000	0.000	0.020	0.491	0.050	0.877
DOG3	0.000	0.000	0.012	0.324	0.034	0.709
SRout	0.010	0.237	0.009	0.244	0.010	0.229
RMDP	0.158	0.985	0.141	0.975	0.129	0.935
PCout	0.121	0.961	0.116	0.961	0.123	0.962

average True Positive Rate (TPR), representing the proportion of true outliers correctly identified; the average False Positive Rate (FPR), representing the proportion of non-outliers incorrectly flagged as outliers; and the Family-Wise False Positive rate (FWFP), which is the proportion of simulation replicates containing at least one incorrect non-outlier flagging.

The results in Table 1 demonstrate that our proposed DOD and DOG methods outperform the competitors by providing a superior balance between detection power and error control. Across nearly all settings, our method DOD achieves a perfect or near-perfect TPR of 1.000, successfully identifying all true outliers. Crucially, our methods accomplish this detection power while maintaining an FPR at or very near zero, indicating robust control over false discoveries. A critical distinction is observed between the two rotation-based methods DOD2 and DOD3, particularly in their control of FWFP. DOD3 aims to control the FWER by its design, and the results confirm its success; DOD3 consistently yields a substantially lower FWFP than DOD2, without compromising its TPR. In contrast, while the competing methods (SRout, RMDP, and PCout) can also exhibit high TPR, they do so at the cost of inflated error rates. Their FPR is consistently higher, and their FWFP frequently exceeds 0.5 and often approaches 1.0, implying that they incorrectly flag non-outliers in the majority of replicates. Moreover, even when outliers were subtle with outlyingness magnitude $(s_\mu, s_\sigma) = (0.25, 0.25)$, DOD1–3 maintained TPR at perfect or near 1.000, while SRout and RMDP failed completely, and PCout’s detection power was diminished.

Table 2 presents the simulation results where no outliers are present in the data ($n_{\text{out}} = 0$). In this setting, the ideal method should refrain from declaring outliers, thus achieving FPR or FWFP below the prespecified maximum FPR α . The results demonstrate the superiority of our proposed methods in controlling error. Our methods exhibit outstanding performance, maintaining a nearly perfect FPR of 0.000 under both ID and AR structures, thus making almost no incorrect outlier flagging. In addition, consistently with tuning parameters, DOD2 and DOG2 control FPR below its prespecified level $\alpha = 0.05$, and DOD3 and DOG3 control FWFP around or below its prespecified level $\alpha = 0.7$. These results empirically confirm that our FWER-controlling algorithms (DOD3, DOG3) provide reliable control over family-wise false positives, making them suitable for robust error control. In contrast, RMDP and PCout perform poorly with FWFP consistently above 0.9, indicating that they incorrectly declare outliers in almost every single replicate.

5 Real data application

5.1 Microarray gene expression

To further evaluate the empirical performance of our proposed methods against competing methods, we analyze the lymphoma microarray gene expression dataset (Dettling 2004). Described in Alizadeh et al. (2000), the dataset contains expression measurements of $p = 4026$ genes for $n = 62$ samples. The samples belong to three lymphoma types, where the largest class, Diffuse Large B-Cell Lymphoma, consists of 42 samples, which will be designated as inliers. The remaining 20 samples from the other two classes will serve as a pool of potential outliers.

Specifically, we designed two experimental scenarios to assess the methods' performance:

1. Contaminated case with $n_{\text{out}} = 2$: In each replication, the dataset was constructed using all 42 inlier samples and 2 outlier samples randomly drawn from the pool of 20. This scenario tests the methods' ability to correctly identify true outliers while

Table 3: Summary of microarray gene expression analysis under $n_{\text{out}} = 2$.

Method	TPR	FPR	FWFP
DOD1	1.000	0.020	0.320
DOD2	1.000	0.026	0.990
DOD3	1.000	0.000	0.000
DOG1	0.907	0.061	0.465
DOG2	0.823	0.000	0.005
DOG3	0.700	0.000	0.000
SRout	1.000	0.081	1.000
RMDP	0.880	0.017	0.725
PCout	0.178	0.137	1.000

Table 4: Summary of microarray gene expression analysis under $n_{\text{out}} = 0$.

Method	FPR	FWFP
DOD1	0.095	1.000
DOD2	0.024	1.000
DOD3	0.000	0.015
DOG1	0.000	0.000
DOG2	0.024	1.000
DOG3	0.000	0.000
SRout	0.133	1.000
RMDP	0.017	0.720
PCout	0.143	1.000

avoiding false positives. We performed 200 replications.

2. Null case with $n_{\text{out}} = 0$: This dataset consisted solely of the 42 inlier samples. This scenario is designed to evaluate the methods' control over the false positive rate when no true outliers are present.

The performance of each method was measured using the average True Positive Rate (TPR), False Positive Rate (FPR), and Family-Wise False Positive Rate (FWFP), which were defined the same as in Section 4. Our proposed procedures, DOD1–3 based on $t_i^{(D)}$ and DOG1–3 based on $t_i^{(G)}$, were implemented with parameters $B = 300$ for the random rotation algorithms, and α same as in Section 4. The competing methods SRout, RMDP, and PCout were implemented again using their default parameters as specified by the original authors.

The summarized results for both scenarios are presented in Tables 3 and 4. In the contaminated case with $n_{\text{out}} = 2$, as shown in Table 3, our proposed methods demonstrated

excellent detection power. The key distinction among the methods lies in their control of false positives. DOD1–3 performed exceptionally well, combining perfect TPR with very low FPR, with DOD3 in particular maintaining a perfect record of zero false positives ($\text{FPR} = 0$ and $\text{FWFP} = 0$). In addition, DOG1–3 achieved better balances between TPR and FPR than the competing methods. In contrast, while a competing method SRout showed a perfect TPR, it came at the cost of a FWFP of 1.0, suggesting it always flags inliers. PCout performed poorly in terms of both detection power ($\text{TPR} = 0.178$) and false positive control ($\text{FWFP} = 1.0$).

In the null case with $n_{\text{out}} = 0$, as shown in Table 4, the superiority of the our proposed methods was evident. DOG1 and DOG3 exhibited perfect control over false positives, achieving a perfect FWFP of 0. DOD3 performed robustly with a near-zero FWFP of 0.015. The remaining methods, including our DOD1, DOD2, DOG2 and all three competitors, struggled significantly in this scenario, with FWFP values ranging from 0.72 to a worst 1.0. This shows that these methods are prone to flagging outliers even when none exist.

5.2 Human face image

As a second real data application, we analyze the Olivetti Research Laboratory face image dataset (Samaria & Harter 1994). This dataset comprises 400 grayscale images of 40 distinct individuals, with 10 different images per person capturing various facial expressions and lighting conditions. Each image consists of 112×92 pixels, resulting in a high-dimensional feature with $p = 10304$ variables.

The experimental design was structured as follows. For each of the 40 individuals, their 10 images were designated as the inlier group. The remaining 390 images from the other 39 individuals served as a pool of potential outliers. Specifically, we investigated two scenarios:

1. Contaminated case with $n_{\text{out}} = 1$: The dataset was constructed with 10 inlier images from one individual, and one outlier image randomly selected from the pool of 390.

Table 5: Summary of human face image analysis under $n_{\text{out}} = 1$.

Method	TPR	FPR	FWFP
DOD1	0.970	0.019	0.125
DOD2	0.975	0.088	0.330
DOD3	0.975	0.086	0.315
DOG1	0.735	0.038	0.205
DOG2	0.130	0.062	0.195
DOG3	0.180	0.062	0.200
SRout	0.980	0.118	0.785
RMDP	0.905	0.060	0.370
PCout	0.845	0.231	0.885

Table 6: Summary of human face image analysis under $n_{\text{out}} = 0$.

Method	FPR	FWFP
DOD1	0.132	0.750
DOD2	0.260	0.950
DOD3	0.264	0.945
DOG1	0.132	0.600
DOG2	0.158	0.635
DOG3	0.170	0.675
SRout	0.178	0.905
RMDP	0.065	0.350
PCout	0.283	0.900

2. Null case with $n_{\text{out}} = 0$: The dataset consisted solely of the 10 inlier images from one individual.

This entire process was repeated for each of the 40 individuals, and for each individual, the experiment was replicated 5 times, leading to a total of 200 independent runs for each scenario. Performance was evaluated using the average TPR, FPR, and FWFP, as defined in Section 4. For this experiment, α for Algorithm 2 was set to 0.1 to reflect the small sample size $n = 10$ of the inlier group.

The summarized results are presented in Tables 5 and 6. In the contaminated case with $n_{\text{out}} = 1$, the results presented in Table 5 show that our proposed methods DOD1–3 and the competing method SRout exhibited the highest detection power with TPRs around 0.975. Among the top-performing methods, DOD1–3 provided better FPR and FWFP control than SRout. RMDP ranked just below this group, showing high detection power and reasonable

FPR and FWFP control.

In the null case with $n_{\text{out}} = 0$, where the focus is on controlling false discoveries, Table 6 shows that RMDP achieved the best performance with the lowest FPR and FWFP values. Our proposed methods DOD1 and DOG1–3 also performed reasonably well. The other procedures struggled to control false positives, yielding FWFP values 0.9 or higher.

6 Conclusion

In this paper, we proposed two statistics for outlier detection in high-dimensional data. These statistics leverage pairwise distances and inner products to capture an observation’s relational dissimilarities. We provided a theoretical foundation for these statistics, demonstrating that as the dimension increases, a non-vanishing margin asymptotically separates outliers from non-outliers. Based on this theoretical guarantee, we developed three practical detection procedures: a clustering-based method and two non-parametric tests based on random rotation, one of which offers robust control over the family-wise error rate. Our simulation studies and real data applications demonstrated that the proposed methods achieve a balance of high detection power and stringent control over false discoveries.

Several avenues for future research remain. First, while our work establishes the asymptotic properties of the statistics, an investigation into their finite dimension behavior under less restrictive assumptions would be a valuable theoretical extension. Second, the current framework is presented based on Euclidean distances and dot products; it could be extended to incorporate other dissimilarity metrics to handle a wider range of data types and outlier mechanisms. Finally, while the random rotation tests are powerful, they may be computationally intensive. Developing faster, deterministic approximations or exploring more computationally efficient resampling schemes could enhance the practical applicability of our methods for massive datasets.

Supplementary material

The supplementary material includes all of the technical details.

Data availability

The lymphoma microarray gene expression dataset analyzed in Section 5.1 is available at the R package `sp1s`, and the Olivetti Research Laboratory face image dataset analyzed in Section 5.2 is available at the following URL: <https://www.kaggle.com/code/serkanpeldek/face-recognition-on-olivetti-dataset>.

Acknowledgments

This work was supported by the 2024 Research Fund of the University of Seoul for Seong-ho Lee, and the National Research Foundation of Korea (NRF) grant (No. RS-2025-16069168) and the IITP-ICAN grant (No. RS-2023-00259934) funded by the Korea government (MSIT) for Yongho Jeon.

References

- Ahn, J., Lee, M. H. & Yoon, Y. J. (2012), ‘Clustering high dimension, low sample size data using the maximal data piling distance’, *Statistica Sinica* pp. 443–464.
- Ahn, J. & Marron, J. (2010), ‘The maximal data piling direction for discrimination’, *Biometrika* **97**(1), 254–259.
- Alizadeh, A. A., Elsen, M. B., Davis, R. E., Ma, C. et al. (2000), ‘Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling’, *Nature* **403**(6769), 503.

- Blaser, R. & Fryzlewicz, P. (2016), ‘Random rotation ensembles’, *The Journal of Machine Learning Research* **17**(1), 126–151.
- Breunig, M. M., Kriegel, H.-P., Ng, R. T. & Sander, J. (2000), Lof: identifying density-based local outliers, *in* ‘ACM sigmod record’, Vol. 29, ACM, pp. 93–104.
- Chandola, V., Banerjee, A. & Kumar, V. (2009), ‘Anomaly detection: A survey’, *ACM computing surveys (CSUR)* **41**(3), 15.
- Chung, H. C. & Ahn, J. (2021), ‘Subspace rotations for high-dimensional outlier detection’, *Journal of Multivariate Analysis* **183**, 104713.
- Dettling, M. (2004), ‘Bagboosting for tumor classification with gene expression data’, *Bioinformatics* **20**(18), 3583–3593.
- Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. (1996), A density-based algorithm for discovering clusters in large spatial databases with noise., *in* ‘Kdd’, Vol. 96, pp. 226–231.
- Filzmoser, P., Maronna, R. & Werner, M. (2008), ‘Outlier identification in high dimensions’, *Computational Statistics & Data Analysis* **52**(3), 1694–1711.
- Hall, P., Marron, J. & Neeman, A. (2005), ‘Geometric representation of high dimension, low sample size data’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(3), 427–444.
- Kriegel, H.-P., Zimek, A. et al. (2008), Angle-based outlier detection in high-dimensional data, *in* ‘Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining’, ACM, pp. 444–452.
- McGill, R., Tukey, J. W. & Larsen, W. A. (1978), ‘Variations of box plots’, *The American Statistician* **32**(1), 12–16.

- Papadimitriou, S., Kitagawa, H., Gibbons, P. B. & Faloutsos, C. (2003), Loci: Fast outlier detection using the local correlation integral, *in* ‘Data Engineering, 2003. Proceedings. 19th International Conference on’, IEEE, pp. 315–326.
- Ro, K., Zou, C., Wang, Z. & Yin, G. (2015), ‘Outlier detection for high-dimensional data’, *Biometrika* **102**(3), 589–599.
- Samaria, F. S. & Harter, A. C. (1994), Parameterisation of a stochastic model for human face identification, *in* ‘Proceedings of 1994 IEEE workshop on applications of computer vision’, IEEE, pp. 138–142.
- Terada, Y. (2013), ‘Clustering for high-dimension, low-sample size data using distance vectors’, *arXiv preprint arXiv:1312.3386*.
- Ye, N. & Chen, Q. (2001), ‘An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems’, *Quality and Reliability Engineering International* **17**(2), 105–112.
- Zimek, A., Schubert, E. & Kriegel, H.-P. (2012), ‘A survey on unsupervised outlier detection in high-dimensional numerical data’, *Statistical Analysis and Data Mining: The ASA Data Science Journal* **5**(5), 363–387.

Supplementary material

S.1 Proof of Theorem 1

We provide the proof for $t_i^{(D)}$; the argument for $t_i^{(G)}$ is analogous. The proof proceeds in two main steps. First, we establish the asymptotic limits of the component terms $\delta_{ij}^{(D)}$ and the column-wise medians $\tilde{\delta}_{\cdot j}^{(D)}$. Second, we use these limits to prove parts (i) and (ii) of the theorem.

We begin by analyzing the probabilistic limit of $\frac{1}{\sqrt{p}}\delta_{ij}^{(D)}$. By definition,

$$\frac{1}{p} \left\{ \delta_{ij}^{(D)} \right\}^2 = \sum_{k \neq i, j} \left(\frac{[\mathbf{D}]_{i,k} - [\mathbf{D}]_{j,k}}{\sqrt{p}} \right)^2.$$

We consider the limit of this sum based on the nature of $\mathbf{x}_i$ and $\mathbf{x}_j$, applying the results from Lemma 1.

1. **If $\mathbf{x}_i, \mathbf{x}_j \sim \mathbf{x}^{(I)}$ (both non-outliers):** For any third point $\mathbf{x}_k$, the term $([\mathbf{D}]_{i,k} - [\mathbf{D}]_{j,k})/\sqrt{p}$ converges to zero in probability by Lemma 1, regardless of whether $\mathbf{x}_k$ is a non-outlier or an outlier. Thus, every term in the sum converges to zero, which implies $\frac{1}{\sqrt{p}}\delta_{ij}^{(D)} \xrightarrow{p} 0$.
2. **If $\mathbf{x}_i, \mathbf{x}_j \sim \mathbf{x}^{(O)}$ (both outliers):** For any third point $\mathbf{x}_k$, the term $([\mathbf{D}]_{i,k} - [\mathbf{D}]_{j,k})/\sqrt{p}$ converges to zero by Lemma 1, as the distances from two outliers to any third point are asymptotically equivalent. This leads to $\frac{1}{\sqrt{p}}\delta_{ij}^{(D)} \xrightarrow{p} 0$.
3. **If $\mathbf{x}_i \sim \mathbf{x}^{(I)}, \mathbf{x}_j \sim \mathbf{x}^{(O)}$ (one non-outlier, one outlier):** The set of $\mathbf{x}_k$ consists of $(n - n_{\text{out}} - 1)$ non-outliers excluding $\mathbf{x}_i$, and $(n_{\text{out}} - 1)$ outliers excluding $\mathbf{x}_j$. The sum of squared limits is then $(n - n_{\text{out}} - 1)\alpha_D^2 + (n_{\text{out}} - 1)\beta_D^2 = \gamma_D^2$ by Lemma 1. It follows that $\frac{1}{\sqrt{p}}\delta_{ij}^{(D)} \xrightarrow{p} \gamma_D$.

Next, we determine the limit of the scaled column-wise median, $\frac{1}{\sqrt{p}}\tilde{\delta}_{\cdot j}^{(D)}$. Since we can assume $n_{\text{out}} < n/2$ by the definition of outlier, the median is determined by the behavior of

the non-outlier rows.

1. **If $\mathbf{x}_j \sim \mathbf{x}^{(I)}$:** The j -th column of $\mathbf{\Delta}_D$ consists of a majority of δ_{ij} values where $\mathbf{x}_i \sim \mathbf{x}^{(I)}$ whose scaled limit is 0, and a minority where $\mathbf{x}_i \sim \mathbf{x}^{(O)}$ whose scaled limit is γ_D . Therefore, $\frac{1}{\sqrt{p}}\tilde{\delta}_{\cdot j}^{(D)} \xrightarrow{p} 0$.
2. **If $\mathbf{x}_j \sim \mathbf{x}^{(O)}$:** The j -th column consists of a majority of values whose scaled limit is γ_D and a minority whose scaled limit is 0. Therefore, $\frac{1}{\sqrt{p}}\tilde{\delta}_{\cdot j}^{(D)} \xrightarrow{p} \gamma_D$.

With these component limits, we now prove the main statements.

- (i) **If $\mathbf{x}_i \sim \mathbf{x}^{(I)}$:** We examine the limit of $\frac{1}{p}\{t_i^{(D)}\}^2 = \sum_{j=1}^n \left\{ \frac{\delta_{ij}^{(D)}}{\sqrt{p}} - \frac{\tilde{\delta}_{\cdot j}^{(D)}}{\sqrt{p}} \right\}^2$.

- For terms where $\mathbf{x}_j \sim \mathbf{x}^{(I)}$, the squared difference converges to $(0 - 0)^2 = 0$.
- For terms where $\mathbf{x}_j \sim \mathbf{x}^{(O)}$, the squared difference converges to $(\gamma_D - \gamma_D)^2 = 0$.

Since every term in the sum converges to zero, $\frac{1}{p}\{t_i^{(D)}\}^2 \xrightarrow{p} 0$, which implies $\frac{1}{\sqrt{p}}t_i^{(D)} \xrightarrow{p} 0$.

- (ii) **If $\mathbf{x}_i \sim \mathbf{x}^{(O)}$:** We analyze the same sum $\frac{1}{p}\{t_i^{(D)}\}^2 = \sum_{j=1}^n \left\{ \frac{\delta_{ij}^{(D)}}{\sqrt{p}} - \frac{\tilde{\delta}_{\cdot j}^{(D)}}{\sqrt{p}} \right\}^2$.

- For the $(n - n_{\text{out}})$ terms where $\mathbf{x}_j \sim \mathbf{x}^{(I)}$, the squared difference converges to $(\gamma_D - 0)^2 = \gamma_D^2$.
- For the n_{out} terms where $\mathbf{x}_j \sim \mathbf{x}^{(O)}$, the squared difference converges to $(0 - \gamma_D)^2 = \gamma_D^2$.

Every one of the n terms in the sum converges to γ_D^2 . Therefore, the sum of the limits is $n\gamma_D^2$:

$$\frac{1}{p}\{t_i^{(D)}\}^2 = \sum_{j=1}^n \left\{ \frac{\delta_{ij}^{(D)}}{\sqrt{p}} - \frac{\tilde{\delta}_{\cdot j}^{(D)}}{\sqrt{p}} \right\}^2 \xrightarrow{p} n\gamma_D^2.$$

Diving both sides by n , we obtain the final result:

$$\frac{1}{np}(t_i^{(D)})^2 \xrightarrow{p} \gamma_D^2 \implies \frac{1}{\sqrt{pn}}t_i^{(D)} \xrightarrow{p} \gamma_D.$$

This completes the proof. ■

S.2 Proof of Theorem 2

We provide the proof for the distance-based statistic $t_i^{(D)}$; the proof for $t_i^{(G)}$ follows analogously. The proof consists of establishing the limits for the maximum of the scaled non-outlier statistics and the minimum of the scaled outlier statistics separately, and then combining them. Let $M_p^{(\mathcal{I})} = \max_{i \in \mathcal{I}} \frac{t_i^{(D)}}{\sqrt{pn}}$ and $m_p^{(\mathcal{O})} = \min_{i \in \mathcal{O}} \frac{t_i^{(D)}}{\sqrt{pn}}$.

First, we show that $M_p^{(\mathcal{I})} \xrightarrow{p} 0$. From Theorem 1(i), we know that for any individual non-outlier $i \in \mathcal{I}$, $\frac{t_i^{(D)}}{\sqrt{p}} \xrightarrow{p} 0$. Rescaling this term gives:

$$\frac{t_i^{(D)}}{\sqrt{pn}} = \frac{1}{\sqrt{n}} \left(\frac{t_i^{(D)}}{\sqrt{p}} \right) \xrightarrow{p} 0.$$

To show that the maximum also converges to zero, we use the union bound for any $\epsilon > 0$:

$$\Pr\{M_p^{(\mathcal{I})} \geq \epsilon\} = \Pr\left[\bigcup_{i \in \mathcal{I}} \left\{ \frac{t_i^{(D)}}{\sqrt{pn}} \geq \epsilon \right\}\right] \leq \sum_{i \in \mathcal{I}} \Pr\left\{ \frac{t_i^{(D)}}{\sqrt{pn}} \geq \epsilon \right\}.$$

Since the number of non-outliers, $|\mathcal{I}|$, is a fixed finite number and each term in the sum converges to 0 as $p \rightarrow \infty$, their sum also converges to 0. Thus, $M_p^{(\mathcal{I})} \xrightarrow{p} 0$.

Next, we show that $m_p^{(\mathcal{O})} \xrightarrow{p} \gamma_D$. From Theorem 1(ii), for any individual outlier $i \in \mathcal{O}$, we have $\frac{t_i^{(D)}}{\sqrt{pn}} \xrightarrow{p} \gamma_D$. To show that the minimum converges to the same limit, we consider for any $\epsilon > 0$:

$$\Pr\{|m_p^{(\mathcal{O})} - \gamma_D| \geq \epsilon\} \leq \Pr\{m_p^{(\mathcal{O})} \geq \gamma_D + \epsilon\} + \Pr\{m_p^{(\mathcal{O})} \leq \gamma_D - \epsilon\}.$$

The first term $\Pr\{m_p^{(\mathcal{O})} \geq \gamma_D + \epsilon\}$ is less than or equal to $\Pr\left\{ \frac{t_i^{(D)}}{\sqrt{pn}} \geq \gamma_D + \epsilon \right\}$ for any single $j \in \mathcal{O}$, which converges to 0. For the second term, we again use the union bound:

$$\Pr\{m_p^{(\mathcal{O})} \leq \gamma_D - \epsilon\} = \Pr\left[\bigcup_{i \in \mathcal{O}} \left\{ \frac{t_i^{(D)}}{\sqrt{pn}} \leq \gamma_D - \epsilon \right\}\right] \leq \sum_{i \in \mathcal{O}} \Pr\left\{ \frac{t_i^{(D)}}{\sqrt{pn}} \leq \gamma_D - \epsilon \right\}.$$

Since $|\mathcal{O}|$ is a fixed finite number and each term in the sum converges to 0, the sum converges

to 0. Therefore, $\Pr\{|m_p^{(\mathcal{O})} - \gamma_{\mathcal{D}}| \geq \epsilon\} \rightarrow 0$, which proves $m_p^{(\mathcal{O})} \xrightarrow{p} \gamma_{\mathcal{D}}$.

Combining the results, we get:

$$m_p^{(\mathcal{O})} - M_p^{(\mathcal{I})} = \min_{i \in \mathcal{O}} \frac{t_i^{(\mathcal{D})}}{\sqrt{pn}} - \max_{i \in \mathcal{I}} \frac{t_i^{(\mathcal{D})}}{\sqrt{pn}} \xrightarrow{p} \gamma_{\mathcal{D}} - 0 = \gamma_{\mathcal{D}}.$$

This completes the proof. ■