JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2024, VOL. 33, NO. 1, 219–230
https://doi.org/10.1080/10618600.2023.2219708

Statistical Significance of Clustering with Multidimensional Scaling

Hui Shen[a], Shankar Bhamidi[a], and Yufeng Liu[b]

[a]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC; [b]Department of Statistics and Operations Research, Department of Genetics, and Department of Biostatistics, Carolina Center for Genome Sciences, Linberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC

ABSTRACT

Clustering is a fundamental tool for exploratory data analysis. One central problem in clustering is deciding if the clusters discovered by clustering methods are reliable as opposed to being artifacts of natural sampling variation. Statistical significance of clustering (SigClust) is a recently developed cluster evaluation tool for high-dimension, low-sample size data. Despite its successful application to many scientific problems, there are cases where the original SigClust may not work well. Furthermore, for specific applications, researchers may not have access to the original data and only have the dissimilarity matrix. In this case, clustering is still a valuable exploratory tool, but the original SigClust is not applicable. To address these issues, we propose a new SigClust method using multidimensional scaling (MDS). The underlying idea behind MDS-based SigClust is that one can achieve low-dimensional representations of the original data via MDS using only the dissimilarity matrix and then apply SigClust on the low-dimensional MDS space. The proposed MDS-based SigClust can circumvent the challenge of parameter estimation of the original method in high-dimensional spaces while keeping the essential clustering structure in the MDS space. Both simulations and real data applications demonstrate that the proposed method works remarkably well for assessing the statistical significance of clustering. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2022
Accepted May 2023

KEYWORDS
Cluster index; Dimension reduction; High-dimension low-sample size data; Principal component analysis; Unsupervised learning

1. Introduction

Clustering is a typical form of unsupervised learning that aims to divide data into several groups so that data points within the same group are more similar than those across groups. Traditional clustering methods use datasets without responses. Clustering is an essential tool for researchers to find potentially helpful hidden structures in high-dimensional data and is commonly used for explanatory data analysis. It has been widely applied in many fields, such as biomedical research, genetics, and social network analysis.

Many clustering methods have been proposed and well-studied in the literature. Comprehensive reviews of clustering algorithms can be found in Xu and Tian (2015) and references therein. Concrete examples of classical clustering algorithms include partition-based algorithms such as K-means (MacQueen et al. 1967), various hierarchical algorithms, and model-based algorithms. Other popular clustering approaches include kernel-based algorithms (Ben-Hur et al. 2001), spectral clustering algorithms (Von Luxburg 2007), and ensemble-based algorithms (Fred and Jain 2005).

Despite rapid developments of clustering algorithms and their wide applications in practice, a natural question is how to assess the statistical significance of clustering results. For a specific clustering algorithm, given the desired number of clusters k, one can typically separate the data into k groups. However, this may result in spurious clusters even in simple settings. For example, with $k = 2$ as explained in Liu et al. (2008), a two-sample t-test gives a significant p-value when we separate data generated from a one-dimensional standard Gaussian distribution into two clusters, suggesting that the two clusters are different from each other. However, in many applications, one may prefer not to divide data from a single Gaussian distribution into multiple clusters. Therefore, assessing the significance of clustering is different from testing subgroup differences.

Several cluster evaluation methods have been proposed in the literature to test the statistical significance of clustering. Among existing methods, the Gaussian cluster definition is commonly used in the sense that data should not be divided further by clustering if they follow a single Gaussian distribution. McShane et al. (2002) proposed a method to evaluate whether the data come from a single Gaussian distribution, that is, whether one should perform clustering on the data. Their method is based on examining the Euclidean distance between samples in a three-dimensional principal component space. Maitra, Melnykov, and Lahiri (2012) used a bootstrap approach and compared a simpler model with a more complicated one for assessing significance of clustering. Chakravarti, Balakrishnan, and Wasserman (2019) tests whether a mixture of Gaussian distributions provides a better fit relative to a single Gaussian distribution focusing

CONTACT Yufeng Liu yfliu@email.unc.edu Department of Statistics and Operations Research, Department of Genetics, and Department of Biostatistics, Carolina Center for Genome Sciences, Linberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

on the low-dimensional setting. Despite progress in this area, assessing the statistical significance of clustering remains an open question, especially in the high-dimension, low-sample size (HDLSS) setting.

Liu et al. (2008) proposed a Monte Carlo-based method called the statistical significance of clustering (SigClust), which addresses the problem of assessing the significance of clustering for HDLSS datasets. To make the HDLSS setting tractable, they used the Gaussian cluster definition and focused on testing whether data come from a single Gaussian distribution. A similar model assumption was used in McLachlan and Peel (2000) and Fraley and Raftery (2002). One critical step in Liu et al. (2008) is to estimate the Gaussian distribution under the null hypothesis. They assumed a factor model to simplify the eigenvalue estimation of the null covariance matrix. Huang et al. (2015) improved the original SigClust by proposing a soft thresholding estimator of the null covariance matrix. Kimes et al. (2017) extended SigClust in the context of hierarchical clustering.

SigClust has been widely applied in practice, such as assessing significant cancer subtypes (TCGA 2012; Agrawal et al. 2014). Despite these successful applications, there are still cases where the original SigClust is not applicable. In particular, in several applications such as natural language processing (NLP) (Poland and Zeugmann 2006), one may only have the pairwise dissimilarity matrix between samples. In that case, clustering can still be performed, but the current SigClust cannot be implemented due to the lack of original data. Furthermore, although Huang et al. (2015) proposed an improved estimator for the null covariance matrix, parameter estimation in the general high-dimensional setting remains a challenging problem. Hence, there is room for further improvement. As shown in Chakravarti, Balakrishnan, and Wasserman (2019), there are certain regions of the parameter space where the original SigClust has relatively low power.

This article proposes a new multidimensional scaling (MDS) based SigClust method. MDS is an important dimension reduction technique with broad applications (Borg and Groenen 2005). The basic idea of MDS is to find low-dimensional representations of the original data while preserving pairwise dissimilarities between samples. This idea aligns well with the goal of clustering since many clustering methods are based on pairwise dissimilarities between samples. Moreover, MDS does not require access to the original data. As mentioned earlier, in many applications, data analysts may not have the original data available such as applications in NLP (Nakamura 2006) and can only work with the pairwise dissimilarity matrix. However, one can still perform effective clustering using only the dissimilarity matrix (Poland and Zeugmann 2006). A natural follow-up question is to understand the statistical significance of the obtained clustering results. The existing SigClust is not applicable due to the lack of original data. In such cases, MDS can provide a natural solution to address these challenges. By using the low-dimensional MDS space, the need for estimating the covariance matrix in a high-dimensional setting, as required for the original SigClust, can be avoided. Based on these considerations, it is meaningful to combine MDS and SigClust to produce an effective clustering evaluation method.

Besides the base version of MDS-based SigClust we mentioned earlier, to tackle the settings mentioned in Chakravarti, Balakrishnan, and Wasserman (2019) where the original SigClust fails, we further improve our MDS-based SigClust using column-wise testing of the MDS embeddings. When the data consist of more than two clusters, besides the significance of clustering, we are also interested in evaluating the number of clusters in the data. To this end, a generalized MDS-based SigClust (see Section 2.5) is introduced using a set of general cluster indices $CI_2, \ldots, CI_K$ for a prespecified K.

The rest of this article is organized as follows. In Section 2, we introduce notation and describe details of different versions of MDS-based SigClust. In Section 3, we examine the theoretical properties under the null and alternative hypotheses. In Section 4, we perform simulation studies to demonstrate the performance of our new methods. We then apply our techniques to real datasets, including cancer gene expression datasets and applications in natural language processing, in Section 5. Finally, we conclude the article with some discussion in Section 6. Proofs of our theoretical results and additional numerical results are provided in the supplementary materials.

2. Methodology

We begin by introducing the notation used throughout the article, as well as MDS and the original SigClust, to establish the basis of our approach. We then proceed to describe our MDS-based SigClust and its implementation.

2.1. Notation

We use regular letters for scalars and bold letters for both matrices and vectors. We use $\mathbf{x}$ and $\mathbf{X}$ to denote random vectors and matrices. We write $[n]$ for the set $\{1, 2, \ldots, n\}$. For any vector $\mathbf{v}$, $\|\mathbf{v}\|$ denotes the Euclidean norm and $\|\mathbf{v}\|_\infty = \max_i |\mathbf{v}(i)|$. The set of $n \times r$ matrices with orthonormal columns is denoted by $\mathcal{O}_{n\times r}$. For a matrix $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_k) \in \mathcal{O}_{n\times k}$ and $m \le k$, let $\mathbf{A}_m = (\mathbf{a}_1, \ldots, \mathbf{a}_m) \in \mathcal{O}_{n\times m}$. For a diagonal matrix $\mathbf{A} = \mathrm{diag}(a_1, \ldots, a_n)$, let $\mathbf{A}_m = \mathrm{diag}(a_1, \ldots, a_m)$ be a $m \times m$ diagonal matrix. Let $f, g : \mathbb{N} \to \mathbb{R}_+$ and let c, b be positive constants and n_0 an integer. Then $f(n) = O(g(n))$ if $f(n) \le \mathrm{cg}(n)$ for all $n > n_0$; $f(n) = \Omega(g(n))$ if $f(n) \ge bg(n)$ for all $n > n_0$; $f(n) = o(g(n))$ if $f(n)/g(n) \to 0$ as $n \to \infty$; $f(n) = w(g(n))$ if $f(n)/g(n) \to \infty$ as $n \to \infty$. Fix $n, d \ge 1$. Suppose we have n iid random samples $\{\mathbf{x}_i\}_{i=1}^n \subseteq \mathbb{R}^d$ with $\mathbb{E}(\mathbf{x}_i) = \mathbf{0}$ and $\mathbb{E}\left(\mathbf{x}_i \mathbf{x}_i^T\right) = \mathbf{\Sigma}$. The sample covariance matrix is defined as $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_\mathbf{x})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_\mathbf{x})^T$, where $\widehat{\boldsymbol{\mu}}_\mathbf{x} = \sum_{i=1}^n \mathbf{x}_i/n$. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n\times d}$ be the data matrix.

2.2. Multidimensional Scaling

Regardless of the availability of the original data $\mathbf{X} \in \mathbb{R}^{n\times d}$, suppose we have access to the dissimilarity matrix $\mathbf{D} \in \mathbb{R}^{n\times n} = (d_{ij})_{i,j\in[n]}$, which measures the pairwise distance between samples for some distance metric d. The main objective of MDS is to find a low-dimensional representation of a set of objects $\boldsymbol{Y} \in \mathbb{R}^{n\times r}$ such that the distance between any two points is close to their corresponding dissimilarity as much as possible. We denote

the pairwise distance between points i and j in the MDS space as $\delta_{ij} = \|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2$ and define the error of representation for the pair $\{i, j\}$ as $e_{ij}^2 = \left(d_{ij} - \delta_{ij}\right)^2$. The total error is defined by summing over all distinct pairs, $\sigma_r(\mathbf{Y}) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left(d_{ij} - \delta_{ij}\right)^2$.

One can consider using different error or distance functions, leading to distinct MDS representations (Borg and Groenen 2005). The goal of MDS is to find a matrix $\mathbf{Y}$ to minimize $\sigma_r(\mathbf{Y})$. When the distance metric d is the Euclidean distance ($d_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$), MDS is equivalent to standard PCA, in which case the method is also called Classic MDS (CMDS). However, MDS is much more general than standard PCA and can also perform nonlinear dimension reduction. Denote $\mathbf{B} = -\frac{1}{2}\mathbf{J}\mathbf{D}^{(2)}\mathbf{J}$, where $\mathbf{D}_{i,j}^{(2)} = \mathbf{D}_{i,j}^2$, $\mathbf{J} = \mathbf{I}_n - \mathbf{1}\mathbf{1}^T/n$ is the centering matrix, and $\mathbf{1} \in \mathbb{R}^n$ is a column vector of ones. Consider the SVD decomposition on $\mathbf{B} = \widetilde{\mathbf{P}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{P}}^T$. In CMDS, the solution $\mathbf{Y}$ can be represented as $\mathbf{Y} = \widetilde{\mathbf{P}}_r\widetilde{\mathbf{\Lambda}}_r^{1/2}$, where $\widetilde{\mathbf{P}}_r$ and $\widetilde{\mathbf{\Lambda}}_r$ are first r eigenvectors and eigenvalues of $\mathbf{B}$.

2.3. Problem Formulation

Before introducing our new method, we start with the original SigClust. In the original SigClust, Liu et al. (2008) used the Gaussian cluster definition and considered the hypothesis problem where the null is that the data come from a single d-dimensional Gaussian distribution and the alternative is the data come from a mixture of d-dimensional Gaussian distributions. To solve this problem, they used a test statistic, k-means cluster index CI_k, which is defined as the ratio between the within-class sum of the squared distance to within-class means and the overall sum of the squared distance to the overall mean,

$$CI_k = \frac{\sum_{s=1}^{k} \sum_{j \in C_s} \left\| \mathbf{x}_j - \bar{\mathbf{x}}^{(s)} \right\|^2}{\sum_{j=1}^{n} \left\| \mathbf{x}_j - \bar{\mathbf{x}} \right\|^2}.$$

Here, for $s \in [k]$, C_s denotes the index set of the sth cluster produced by a specific clustering algorithm and $\bar{\mathbf{x}}^{(s)}$ represents the corresponding within-cluster mean. The intuition underlying the cluster index is that if the k clusters produced by some clustering algorithm such as k-means are well-separated, the data points concentrate around the cluster centers within each cluster and the within-cluster sum of the squared distance tends to be small. On the contrary, if the data come from a single Gaussian cluster, and we try to divide them into k parts, the cluster index tends to be large.

Liu et al. (2008) focused on the test statistic CI_2 to test whether there is one or more than one Gaussian cluster. To carry out the test and find the p-value, they used a Monte Carlo procedure which generates Gaussian random variables $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ under the null. However, it is difficult to achieve consistent estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in the HDLSS setting. Since the test statistic CI_2 is location-invariant and the Euclidean distance is invariant to orthogonal rotations, one can assume $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma}$ to be diagonal. This property simplifies the task of estimating $d(d+1)/2$ parameters to estimating d eigenvalues of $\boldsymbol{\Sigma}$. Moreover, Liu et al. (2008) assumed the covariance matrix to be spiked, namely

$$\Sigma = \mathbf{V}\Lambda\mathbf{V}^T, \quad \Lambda = \Lambda_0 + \sigma_n^2\mathbf{I}, \tag{1}$$

where Λ_0 is of low rank, which captures a few strong signals in the data, and the relatively small σ_n^2 represents the constant variance of the background noise. Then it is enough to estimate a few top eigenvalues of Λ_0 and background noise variance σ_n^2. The above assumptions help make the high-dimensional estimation tractable and are reasonable in applications.

The Gaussian cluster definition is a central tenant of the original SigClust (Liu et al. 2008). As will be evident later, we perform a detailed investigation into the relevance of this assumption. In particular, we shall find that the significance of clustering is relatively robust, and under a range of alternative definitions, the Gaussian cluster assumption is the most conservative one. The difficulty of more general notions of a single cluster is that if no specific parametric assumption is made, the exact null distribution of the test statistic CI_k is hard to compute. Our idea is to calculate the p-value using a simple Monte Carlo procedure without explicitly figuring out the cluster index's null distribution. When generating data under the null, the SigClust-based methods (Liu et al. 2008) generate Gaussian random variables as the reference distribution. As we will show later, through theoretical results and simulations, the cluster index CI_k converges under a general class of distributions. The Gaussian distribution as a reference is the conservative choice. The population CI_k under the Gaussian assumption is smaller than that of many other distributions. When the null is not Gaussian, and we are generating Gaussian samples, the p-value tends to be larger than that of the true null hypothesis. As a result, our SigClust tends to make a conservative conclusion. In real applications such as modern gene expression analyses, a fundamental issue is that clusters are sometimes detected and claimed to be real when they may not be significant. Hence, the generation of Gaussians is meaningful because it helps avoid over-clustering when the data just correspond to one cluster.

These observations motivate us to extend the definition of a single Gaussian cluster to a single unimodal cluster, that is, data coming from a single unimodal distribution, such as t or χ^2, and consider a general hypothesis problem:

H_0 : The data come from a single d-dimensional unimodal distribution;

H_1 : The data come from a mixture of d-dimensional unimodal distributions.

For this hypothesis problem, the k-means cluster indices CI_k is still helpful as we explained above. There could be a class of testing statistics CI_k given different values of k. In general, we can choose $k = 2$ and use 2-means cluster index as the test statistic CI_2 when we are interested in testing whether there is one or more than one cluster. In some cases, if the test result is significant and we are further interested in knowing the number of clusters in the data, we can use multiple cluster indices simultaneously. In the next section, we will focus on the 2-means cluster index CI_2 for our new proposed method and discuss a generalized method based on CI_k in Section 2.5.

2.4. MDS-based SigClust

In this paper, we propose a new MDS-based SigClust, which combines the original SigClust and the dimension reduction

technique MDS. The proposed method starts with the dissimilarity matrix $\mathbf{D} \in \mathbb{R}^{n\times n}$ between samples. We can achieve a low-dimensional representation matrix $\mathbf{Y} \in \mathbb{R}^{n\times r}$ through MDS. If the data come from a single cluster or a mixture of clusters, the embedding matrix $\mathbf{Y}$ tends to preserve certain properties of the single or mixture of clusters. Furthermore, for two datasets of the same size, suppose one is obtained from a single cluster and another from a mixture of two distinct clusters with the same covariance matrix as the first dataset. The CI_2 of the second dataset should be smaller than that of the first dataset, that is, the separation information of the mixtures (difference between two mean vectors) can be captured through a smaller CI_2.

There are cases where the original SigClust might fail, see Chakravarti, Balakrishnan, and Wasserman (2019). For simplicity, consider the Gaussian cluster definition. If the data come from a mixture of two Gaussian distributions, the separation can happen in multiple ways, that is, the difference between the means of the two distinct Gaussian components can be nonzero in any coordinates. The mean difference may be nonzero in one coordinate c_1, but the variance is the largest in another coordinate c_2. Therefore, the coordinate c_2 will determine how the data are clustered into two and dominate the cluster index of the data. Even if the cluster index can capture the separation signal in coordinate c_1, the signal in coordinate c_1 only accounts for a small portion of the test statistic CI_2 and is too small to be detected as significant. To address this issue, we improve our method, aiming to capture the separation signal from all possible directions.

In this modified procedure, an initial goal is to calculate a combined CI_2 defined as the minimum of the CI_2's calculated from $\mathbf{Y}$ and each column of $\mathbf{Y}$. However, one challenging issue is that CI_2's calculated from data with different dimensions are not comparable because the limiting distribution of the cluster index of one dataset depends on its dimensionality. To solve this problem, we use the 2-means clustering result as classification labels and project the data $\mathbf{Y}$ onto the one-dimensional space by linear discriminant analysis (LDA). Then the combined CI_2 is taken to be the minimum of the CI_2's calculated from the one-dimensional LDA projection of $\mathbf{Y}$ and each column of $\mathbf{Y}$. Following the Monte Carlo idea of the original SigClust, we estimate the sample covariance matrix $\widehat{\Sigma}_{\mathbf{Y}}$ of $\mathbf{Y}$, generate data $\mathbf{Z}$ from $N(\mathbf{0}, \widehat{\Sigma}_{\mathbf{Y}})$, and calculate a combined CI_2 of the simulated data $\mathbf{Z}$ using the same procedure. After that, we compare the observed CI_2 with those of the simulated data to draw a conclusion about the significance of clustering.

Our base version of MDS-based SigClust is summarized as below.

Base version of MDS-based Sigclust:

Step 1. Choose the dimension r of the MDS space. Obtain the MDS matrix $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_r)$ of the dimension $n \times r$ from the dissimilarity matrix $\mathbf{D}$.

Step 2. Implement the 2-means clustering on $\mathbf{Y}$ and calculate the cluster index CI_2 of $\mathbf{Y}$ using the estimated labels, denoted as $CI_{2,\mathbf{Y}}$.

Step 3. Estimate the sample covariance matrix $\widehat{\Sigma}_{\mathbf{Y}}$ of $\mathbf{Y}$. Generate an $n \times r$ matrix $\mathbf{Z}$ with each row $\mathbf{z}_i$ drawn independently from $N(\mathbf{0}, \widehat{\Sigma}_{\mathbf{Y}})$ for $i \in [n]$.

Step 4. Perform Step 2 on $\mathbf{Z}$ and calculate the cluster index CI_2, denoted as $CI_{2,\mathbf{Z}}$.

Step 5. Repeat Steps 3 and 4 N_{sim} times. For $i \in [N_{sim}]$, $\mathbf{Z}_i$ denotes the ith simulation and $CI_{2,\mathbf{Z}_i}$ denotes the corresponding CI_2. Then we have a set of N_{sim} $CI_{2,\mathbf{Z}}$.

Step 6. Using the empirical distribution of $\{CI_{2,\mathbf{Z}_i} : i \in [N_{sim}]\}$, calculate a p-value for the CI_2 of $\mathbf{Y}$. Draw a conclusion based on a prespecified level of significance α.

For the scenario motivated by Chakravarti, Balakrishnan, and Wasserman (2019), Step 2 in the above algorithm can be modified into the following one.

Modification MDS-based SigClust:
Step 2': For each $i \in [r]$, implement 2-means clustering on $\mathbf{y}_i$ and use the labels to calculate the CI_2 of $\mathbf{y}_i$, denoted as $CI_{2,\mathbf{y}_i}$. Implement the 2-means clustering on $\mathbf{Y}$ and take the clustering labels as the classification labels to apply LDA. Use the LDA result to get the one-dimensional projection of $\mathbf{Y}$, denoted as $\mathbf{Y}_{LDA}$. Calculate the CI_2 of $\mathbf{Y}_{LDA}$, denoted as $CI_{2,LDA}$. The combined CI_2 of $\mathbf{Y}$, denoted as $CI_{2,\mathbf{Y}}$ is taken to be $\min\{\{CI_{2,\mathbf{y}_i}\}_{1\le i\le r}, CI_{2,LDA}\}$.

There are different methods to calculate the p-value in Step 6 of the above procedure. One method is to use the proportion of simulated CI_2's that are smaller than $CI_{\mathbf{Y}}$. This method depends heavily on the number of simulations N_{sim}. Another method is to fit a one-dimensional Gaussian distribution using the simulated CI's and calculate the quantile of $CI_{2,\mathbf{Y}}$ in this fitted distribution. The second approach provides a continuous range of p-values, especially when the empirical p-value is zero. We refer to these two types of p-values as the *percentile p-value* and the *fitted p-value*, respectively.

Note that r is a prespecified parameter representing the dimension of the MDS space. As will be seen below, in many settings, when using the 2-means cluster index CI_2, $r = 1$ or 2 is enough to detect the separation signal if the original data come from a mixture of two or more clusters. This is due to the fact that the first few dimensions can capture the signals as shown in several settings such as Gaussian mixture models and stochastic block models (Abbe et al. 2020; Löffler, Zhang, and Zhou 2021). However, when we want to evaluate the number of clusters using CI_k with $k > 2$ as described in Section 2.5, a higher dimension r would be preferred.

2.5. Generalized MDS-based SigClust

In some cases, when the p-value in the above test procedure is significant, we may be interested in evaluating the number of clusters in the data, which can be summarized as a two-stage testing problem: (a) whether there is one or more than one cluster; (b) if there is more than one cluster, how many clusters exist in the data?

To solve this, we simultaneously consider a sequence of $K-1$ test statistics $CI_2, \ldots, CI_K$, which correspond to the hypothesis test problems:

H_0 : The data come from a single d-dimensional unimodal distribution;

H_1 : The data come from a mixture of k d-dimensional unimodal distributions.

for $k = 2, \ldots, K$. For each CI_k, we can calculate the p-value p_k using a similar procedure for CI_2 described in Section 2.4. Then we obtain a set of $K-1$ p-values $(p_2, \ldots, p_K)$. To deal with the issue of multiple comparisons, we use the Holm–Bonferroni method (Holm 1979), while other adjustment methods can be used as well. If any of the adjusted p-values is significant, we would reject the null that there is only one cluster. To decide how many clusters are preferred, we can estimate the number of clusters by $\text{argmin}_{s\in\{2,\ldots,K\}} p_s$, that is, the hypothesis index that has the minimum p-value. The same idea can apply to the original SigClust, which will be used in simulations and real data examples for comparison (name it generalized SigClust-Soft).

3. Theoretical Properties

To gain further insight into the proposed MDS-based SigClust, we study some of its theoretical properties. For simplicity, we consider 2-means clustering. Assume we have n iid samples $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ from some distribution $\mathbb{P}$. Recall that the sample k-means cluster centers $\mathbf{b_n} = (b_{n1}, \ldots, b_{nk}) \in \mathbb{R}^{d\times k}$ is defined as $\mathbf{b}_n = \arg\min_{\mathbf{a}\in\mathbb{R}^{d\times k}} W_n(\mathbf{a})$, where $W_n(\mathbf{a}) = \frac{1}{n}\sum_{i=1}^n \min_{1\le j\le k} \left\|\mathbf{x}_i - a_j\right\|^2$. One can define the population k-means cluster centers $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k) \in \mathbb{R}^{d\times k}$ as $\boldsymbol{\mu} = \text{argmin}_{\mathbf{a}\in\mathbb{R}^{d\times k}} W_{\mathbb{P}}(\mathbf{a})$, where $W_{\mathbb{P}}(\mathbf{a}) = \mathbb{E}\left[W_n(\mathbf{a})\right]$.

Theorem 3.1 (Convergence of Cluster Index). Assume one-dimensional random variables $x_1, \ldots, x_n$ are independently generated from some distribution $F(\cdot)$ with continuous density function $f(\cdot)$. Suppose the density function is symmetric over 0 and dominated by $\rho(\cdot)$ with $\int_{\mathbb{R}} r\rho(r)dr < \infty$ and assume that $\int_{\mathbb{R}} x^2 f(x) < \infty$. Moreover, suppose the population 2-means centers $\boldsymbol{\mu} = (\mu_1, \mu_2)$ are unique and symmetric. Then for $X \sim F$, we have

$$CI_2 \xrightarrow{a.s.} \frac{E(X^2) - (E|X|)^2}{E(X^2)}.$$

Remark 1. The above theorem can be extended to finite-dimensional settings under similar assumptions.

Remark 2. Most of the assumptions on the distribution F in the above result are to guarantee that the sample 2-means centers converge to the population 2-means centers as sample size $n \to \infty$. Then the main theorem in Pollard et al. (1982) can be applied to show the consistency of the cluster index.

Theorem 3.1 shows that the test statistic, cluster index, is not designed exclusively for Gaussian clusters. Even when data are generated from non-Gaussian distributions, such as t and χ^2 distributions, the cluster index can still converge to a limit. This result provides insight into why our method can still effectively work for non-Gaussian data. We have provided detailed proofs for all of our stated results in the Appendix.

Next, we focus on the specific setting of Gaussian clusters and show further theoretical results about our MDS-based SigClust. Suppose $\mathbf{x}_i$ follows $\frac{1}{2}N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{2}N_d(-\boldsymbol{\mu}, \boldsymbol{\Sigma})$ independently for $i \in [n]$. We are interested in the hypothesis testing problem: $H_0 : \boldsymbol{\mu} = \mathbf{0}$ versus $H_1 : \boldsymbol{\mu} \neq \mathbf{0}$. For simplicity, we use that classical MDS with the Euclidean distance. The first result is on the p-value of MDS-based SigClust under the null hypothesis H_0.

Theorem 3.2. Suppose the data come from $N(\mathbf{0}, \Sigma)$. For $r = 1$, the distribution of p-value from MDS-based SigClust converges to $U[0, 1]$ as $n \to \infty$.

The idea of the proof is to show that the MDS matrix $\mathbf{Y}$ preserves Gaussian properties if the original data $\mathbf{X}$ come from a single Gaussian distribution. The uniform distribution of the p-value on [0,1] shows that MDS-based SigClust can control the Type I error.

Next, we consider the alternative hypothesis H_1. Without loss of generality, we assume that $\boldsymbol{\ell} = (\mathbf{1}_{n_1}^T, -\mathbf{1}_{n_2}^T)^T$ is the n-dimensional label vector with 1 representing the first group and -1 representing the second group. We use n_i to denote the number of observations in the ith group for $i = 1, 2$. Let $\lambda_{\max} = \max_{1\le j\le d} \lambda_j = \lambda_1$, where λ_j's are the eigenvalues of Σ. Define the signal-to-noise ratio as $\text{SNR} = \frac{\|\boldsymbol{\mu}\|^2}{\lambda_{\max}}$, where $\|\boldsymbol{\mu}\|^2$ represents the signal and $\lambda_{\max}$ the noise. The following results show that our method can recover the true class labels with high probability and maintain high power when SNR is sufficiently large. Lemma 3.3 and Corollary 3.4 are modified from Little, Xie, and Sun (2022).

Lemma 3.3. In the general high-dimensional setting where $d = \Omega(n)$, suppose the data come from a mixture of two Gaussian distributions under H_1 with $\|\boldsymbol{\mu}\| \neq 0$. With a probability at least $1 - 4/n$, we have

$$\left\|\widetilde{\mathbf{p}}_1 - \widetilde{\boldsymbol{\ell}}\right\|_\infty \le (\omega^2 + 3\omega/2)/\sqrt{n}, \tag{2}$$

where $\omega = \frac{32}{\|\boldsymbol{\mu}\|}\{8(\lambda_{\max}\log n)^{1/2} + 4(6\log n/d)^{1/2}\lambda_{\max}/\|\boldsymbol{\mu}\| + d\lambda_{\max}/(n\|\boldsymbol{\mu}\|)\}$. Here, $\widetilde{\boldsymbol{\ell}} = \frac{1}{\sqrt{n}}\boldsymbol{\ell} = \frac{1}{\sqrt{n}}(\mathbf{1}_{n_1}^T, -\mathbf{1}_{n_2}^T)^T$ is the normalized true label vector and $\widetilde{\mathbf{p}}_1$ is the first column of $\mathbf{Y}$.

Corollary 3.4. In the high-dimensional setting where $d = O(n\log n)$, suppose the data come from a mixture of two Gaussian distributions under H_1 with $\|\boldsymbol{\mu}\| \neq 0$ and $\frac{\|\boldsymbol{\mu}\|^2}{\lambda_{\max}} = w(\log n)$. Then we have $\left\|\widetilde{\mathbf{p}}_1 - \widetilde{\boldsymbol{\ell}}\right\|_\infty = o(\frac{1}{\sqrt{n}})$.

Note that each element in $\widetilde{\boldsymbol{\ell}}$ takes values in the set $\{\frac{1}{\sqrt{n}}, -\frac{1}{\sqrt{n}}\}$. From this corollary, $\widetilde{\mathbf{p}}_1$ is close to $\widetilde{\boldsymbol{\ell}}$ elementwise, which implies that using the MDS matrix $\mathbf{Y}$ with $r = 1$ can recover the true cluster labels accurately.

Theorem 3.5. In the high-dimensional setting where $d = O(n\log n)$, suppose the data come from a mixture of two Gaussian distributions under H_1 with $\|\boldsymbol{\mu}\| \neq 0$ and $\frac{\|\boldsymbol{\mu}\|^2}{\lambda_{\max}} = w(\log n)$. For $r = 1$, the p-value from MDS-based SigClust converges to 0 in probability as $n \to \infty$.

This theorem tells us that if the data truly come from a mixture of two Gaussian distributions under a moderate dimensional regime and SNR grows faster than $\log n$, MDS-based SigClust can detect the separation and produce a significantly small p-value. A similar conclusion can be drawn in higher dimensional settings as follows.

Theorem 3.6. Consider the general high and ultra high dimensional settings where $d = \Omega(n\log n)$. Suppose the data come from a mixture of two Gaussian distributions under H_1 with

$\|\boldsymbol{\mu}\| \neq 0$ and $\frac{\|\boldsymbol{\mu}\|^2}{\lambda_{\max}} = w(\frac{d}{n})$. For $r = 1$, the p-value from MDS-based SigClust converges to 0 in probability as $n \to \infty$.

4. Simulations

In this section, we compare cluster evaluation methods on various simulated examples in low and high-dimensional settings. Methods include RIFT and MRIFT (Chakravarti, Balakrishnan, and Wasserman 2019), the method proposed in McShane et al. (2002), SigClust using soft thresholding (SigClust-Soft, Huang et al. 2015), our proposed MDS-based SigClust (SigClust-MDS) and SigClust-MDS with the true covariance matrix (SigClust-True-MDS). The SigClust-Soft and SigClust-True-MDS generate Gaussian data under the null in the original space. Our MDS-based SigClust involves the estimation of the sample covariance matrix in a low-dimensional MDS space.

To account for the rotation invariance of CI_k under a single Gaussian and a mixture of Gaussians with identical covariance matrices, we restrict our attention to the case where the covariance matrix Σ for each Gaussian component is diagonal with entries $\lambda_1, \ldots, \lambda_d$. In all experiments, we set $n = 100$, $d = 1000$, and $N_{\text{sim}} = 1000$, unless otherwise specified. We obtain the cluster assignments for the CI_k using k-means clustering with k specified in each section. We use the fitted p-values throughout. Different methods are evaluated based on their ability to maximize power while controlling the Type I error.

In Section 4.1, we generate data from a single Gaussian and a mixture of two Gaussians and compare SigClust methods with the method proposed in McShane et al. (2002). In Sections 4.2, we compare our MDS-based SigClust with RIFT and MRIFT (Chakravarti, Balakrishnan, and Wasserman 2019) in a low-dimensional setting. To demonstrate the performance of our method under cluster definitions other than Gaussian, we generate data from t and Poisson distributions and visualize the results in Section 4.3. The generalized MDS-based SigClust is evaluated in Section 4.4. We summarize the simulation results in Section 4.5. Extended simulations are provided in the Appendix.

4.1. Gaussian Mixtures

To analyze the performance of three SigClust-based methods, we generate data under the null and alternative hypotheses, namely a single Gaussian $N_d(\mathbf{0}, \boldsymbol{\Sigma})$ and a mixture of two distinct Gaussian distributions $\frac{1}{2}N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{2}N_d(-\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We let $\boldsymbol{\mu} = (a, 0, \ldots, 0)^T$ and $\boldsymbol{\Sigma} = \text{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$. The covariance matrix of the data is $\boldsymbol{\Sigma}^* = \text{diag}(\lambda_1 + a^2, \lambda_2, \ldots, \lambda_d)$. SigClust-True-MDS uses $\boldsymbol{\Sigma}^*$ to generate the simulated data $\mathbf{Z}$ in the Monte Carlo procedure, on which we apply SigClust-MDS. We use the modified version of MDS-based SigClust with CI_2 and $r = 2$. Consider three settings for $\boldsymbol{\Sigma}$ as follows:

(a) $\boldsymbol{\Sigma} = \text{diag}(100, 100, \ldots, 100, 1, \ldots, 1)$, where the first 10 entries are 100;

(b) $\boldsymbol{\Sigma} = \text{diag}(10, 10, \ldots, 10, 1, \ldots, 1)$, where the first 100 entries are 10;

(c) $\boldsymbol{\Sigma} = \text{diag}(100, 95, \ldots, 10, 5, 1, \ldots, 1)$, where the first 20 entries form an arithmetic sequence.

The first setting corresponds to the spiked covariance model, with a few large eigenvalues and others small. In the second setting, we assume a group of medium-large eigenvalues together with small ones. The third setting interpolates between the first two, where the eigenvalues decrease gradually.

We plot the empirical distributions of p-values under three settings in Figure 1 (and Figures S1–S2 in the Appendix). Two vertical lines represent two thresholds $\alpha = 0.05$ and 0.1. In each figure, four subfigures show how the empirical distributions change as a gets larger for all methods. For effective tests, we expect to see that the empirical distributions of p-values are close to the diagonal line when $a = 0$ and move toward the upper-left corner quickly as a increases.

When $a = 0$ (a single Gaussian distribution), Figure 1(a) shows that all four methods can control the Type I error. Moreover, SigClust-MDS, SigClust-True-MDS, and McShane et al. (2002) produce uniformly distributed p-values on [0,1] while SigClust-Soft produces large p-values with conservative results. When $a \neq 0$ (Gaussian mixtures), the empirical distributions of all four methods move toward the upper-left corner as a increases except the method in McShane et al. (2002). In all three settings, the power of SigClust-MDS is close to 1 when a is moderately large, while the other methods have power less than 0.5 under $\alpha = 0.05$. Compared with SigClust-MDS, the method by McShane et al. (2002) gains power very slowly. Overall, SigClust-MDS is more powerful than the other methods when the signal is in one coordinate direction.

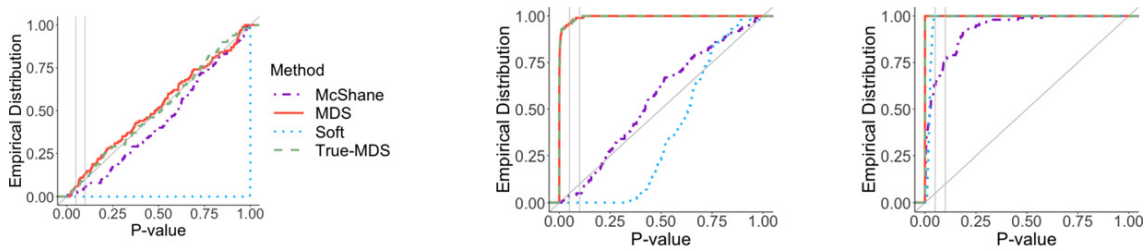


Figure 1. Empirical distributions of SigClust p-values based on True-MDS, soft, and MDS methods and method from McShane et al. (2002) for Setting 2. The mean difference comes from one direction with $a = 0, 6, 10$, respectively.

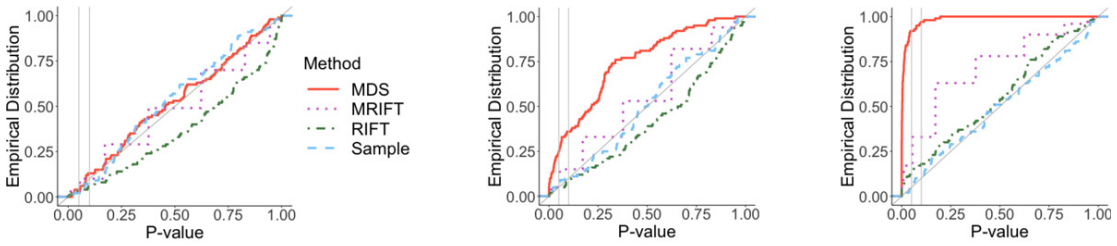


Figure 2. Empirical distributions of *p*-values based on RIFT, MRIFT, SigClust-Sample, and SigClust-MDS in the low-dimensional setting described in Section 4.2. The mean difference comes from the first direction with $a = 0, 2, 3$, respectively.

4.2. Signal in Directions with Low Variation

Chakravarti, Balakrishnan, and Wasserman (2019) pointed out that the original SigClust may have relatively low power against certain alternatives and proposed a test for a relative fit of mixtures focusing on the low-dimensional setting. Given the hypothesis problem and the data, they fit two models, a multivariate Gaussian and a mixture of two multivariate Gaussians, to compare which model fits the data better and get a *p*-value to make a conclusion. They described a simulation setting where the original SigClust has low power. Here, we use their setting to compare our MDS-based SigClust, the original SigClust, RIFT, and MRIFT (Chakravarti, Balakrishnan, and Wasserman 2019). The modified version of MDS-based SigClust is used with CI_2 and $r = 2$.

In this simulation setting, a mixture of two Gaussian distributions $\frac{1}{2}N(\mathbf{0}, \Sigma) + \frac{1}{2}N(\boldsymbol{\mu}, \Sigma)$ is considered, where $\boldsymbol{\mu} = (a, 0, \ldots, 0)$. For simplicity, we let Σ be a diagonal matrix with $\Sigma_{jj} = 400$ for $j = 2$ and $\Sigma_{jj} = 1$ for $j \neq 2$. This problem is challenging because the signal, that is, the mean difference of two Gaussian components, lies in the first dimension, but its variance is significantly smaller than that of the second dimension.

We let $n = 100$ and $d = 5$. For this low-dimensional problem, we use the original SigClust with the sample covariance matrix (SigClust-Sample) since $d = 5 \leq n$. We plot the empirical distributions of *p*-values based on four methods with different values of a in Figure 2. Two vertical lines correspond to $\alpha = 0.05$ and 0.1 as before. For $a = 0$, the ideal distribution of *p*-value is uniform [0,1]. For $a > 0$, we hope to have small *p*-values because there are two clusters. From Figure 2, we can see that for $a = 0$, all methods work similarly except RIFT is slightly more conservative. For $a = 2$ and 3, SigClust-MDS works better than the other methods. In particular, for $a = 3$, our method has power close to 1 while the other methods have low power. This example further demonstrates the usefulness of the modified CI_2, which can capture the separation signal in all possible directions. At the same time, SigClust-Sample may ignore some information when the signal (mean difference) is not in the largest variance direction.

4.3. Sensitivity Analysis

Although our theoretical analyses focus on Gaussian clusters, our method is applicable to cluster definitions other than Gaussian. It is conservative in a number of settings in the sense that if the test is significant, this might indicate strong evidence of underlying clusters. To demonstrate this, we generate data from *t* and Poisson cluster definitions under both null and alternative hypotheses in the high-dimensional setting.

Under the alternative hypothesis, a mixture of two unimodal distributions is generated from the same distribution class with different location parameters $\boldsymbol{\mu}_1 = (a, a, \ldots, a)$ and $\boldsymbol{\mu}_2 = (-a, -a, \ldots, -a)$. Taking the t distribution as an example, under the null hypothesis, each column of data is independently generated from a single t distribution with degrees of freedom being 10, that is, $t(10)$. Under the alternative hypothesis, a mixture of two shifted t distributions $\frac{1}{2}(t(10)+a)+\frac{1}{2}(t(10)-a)$ are generated. For the Poisson case, data are generated similarly with mean 3. We use the modified version of MDS-based SigClust with CI_2 and $r = 2$. As shown in Figure 3, both methods give conservative *p*-values and control the Type-I error well under the null. As the cluster mean difference a gets larger, our SigClust-MDS gains power quickly while Sigclust-soft's power stays low.

4.4. Generalized SigClust

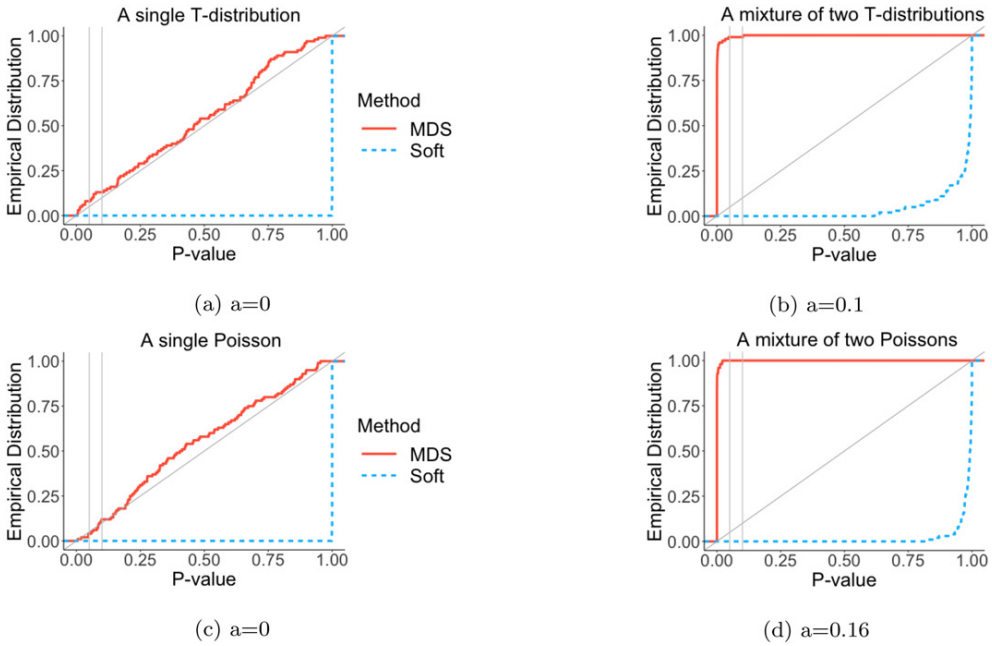
In Section 2.5, we proposed a generalized MDS-based SigClust to identify the number of clusters when there is more than one. To evaluate its performance, we generate data from a single Gaussian and a mixture of multiple Gaussians. When the data are from Gaussian, we want to see whether the proposed generalized method can control Type I error. When multiple clusters exist, we evaluate its performance by two criteria: (a) power: the probability of correctly rejecting the null; (b) selection ratio: the probability of choosing the correct number of clusters K.

Throughout this section, we generate data from Gaussian or a mixture of Gaussians with the identity covariance matrix $\Sigma = \mathbf{I}_d$ for each Gaussian component. We use the base version of MDS-based SigClust with $r = 5$ and consider the set of test statistics $CI_2, \ldots, CI_5$. For a mixture of K Gaussians, we try $K = 2, 3$, and 4. We compare the generalized version of the original SigClust (SigClust with $CI_2, \ldots, CI_5$) and our generalized method proposed in Section 2.5. The cluster centers for different K are:

(a) $K = 2 : \boldsymbol{\mu}_1 = (a, \ldots, a)$ and $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$;

(b) $K = 3 : \boldsymbol{\mu}_1 = (0, \ldots, 0)$, $\boldsymbol{\mu}_2 = (a, \ldots, a)$ and $\boldsymbol{\mu}_3 = (a, \ldots, a, -a, \ldots, -a)$ with first half coordinates being a;

(c) $K = 4 : \boldsymbol{\mu}_1 = (a, \ldots, a)$, $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_3 = (a, \ldots, a, -a, \ldots, -a)$ with first half coordinates being a and $\boldsymbol{\mu}_4 = -\boldsymbol{\mu}_3$.



(a) a=0 (b) a=0.1

(c) a=0 (d) a=0.16

Figure 3. Empirical distributions of *p*-values based on SigClust-Soft and SigClust-MDS in the high-dimensional setting under *t* and Poisson cluster definitions considered in Section 4.3.

Table 1. Performance of the generalized SigClust-Soft and SigClust-MDS.

$K = 1$	SigClust-Soft Type-I error 0		SigClust-MDS Type-I error 0	
	Power	Selection ratio	Power	Selection ratio
$K = 2(a = 3)$	0.68	1	0.94	1
$K = 3(a = 0.16)$	0	NA	1.00	0.97
$K = 4(a = 0.12)$	0	NA	0.98	0.94

NOTE: Type I errors under a single Gaussian and power under a mixture of K Gaussians are given.

Table 1 demonstrates the performance of both methods under different K. For a single Gaussian ($K = 1$), both generalized methods have Type I error being 0. Under the alternatives, Table 1 shows that when the value of a is reasonably large, our generalized SigClust-MDS has high power and selects the correct number of clusters with high frequency, while the generalized SigClust-Soft has very low power.

4.5. Summary of Simulation-based Findings

In Section 3 of the Appendix, we provide additional simulations to demonstrate the performance of our proposed methods. All simulation results in Section 4 and Section 3 of the Appendix show that our MDS-based SigClust methods can control the Type I error under the null, have great power under the alternatives, and are robust to different cluster definitions. Based on these simulation examples, we can see that SigClust-MDS performs the best under both null and alternative hypotheses. In particular, the *p*-values are approximately uniformly distributed on [0,1] under the null, similar to SigClust-MDS-True. It also has the largest power in all settings under the alternative among all comparison methods. Moreover, when the spiked covariance assumption fails in the high-dimensional setting, SigClust-MDS is much more powerful than SigClust-Soft.

5. Real Data Analysis

We demonstrate the effectiveness of our base and generalized versions of SigClust-MDS (see Sections 2.4–2.5) on several cancer gene expression datasets and various applications in natural language processing. Each dataset consists of several subgroups and contains a group label for each sample. We consider two approaches to evaluate the cluster significance. One is to test every pairwise combination of two clusters using the base version of SigClust-MDS with CI_2 and $r = 2$. When calculating the test statistic CI_2, we need to first separate the data into two clusters. We use both the group labels ("True") and 2-means clustering results ("Est") as cluster assignments to calculate the CI_2's. The true labels correspond to underlying (biological) groups of interest, while the estimated labels from clustering algorithms correspond to clusters with good separation between clusters. Clustering errors are typically reported as the misclassification rate of the *k*-means clustering algorithm compared to the true labels. In most cases, the algorithm performs similarly for both choices of labels. The other method is to test all clusters simultaneously using the generalized SigClust-MDS and choose the number of clusters based on the minimum *p*-value. We implement SigClust using soft thresholding (SigClust-Soft) for comparison. The fitted *p*-values are used throughout.

5.1. Multi-Cancer Gene Expression Dataset

We first consider a multi-cancer dataset consisting of three cancer types: 100 samples of head and neck squamous cell carcinoma (HNSC), 100 samples of lung squamous cell carcinoma (LUSC), and 100 samples of lung adenocarcinoma

Table 2. SigClust p-values for each pair of subtypes for the Multi-Cancer data.

	Soft(True)	Soft(Est)	Error(Soft)	MDS(True)	MDS(Est)	Error(MDS)
HNSC & LUSC	8.78e-5	2.07e-4	0.04	2.14e-8	5.38e-08	0.05
HNSC & LUAD	2.70e-18	2.54e-17	0.01	7.89e-47	1.90e-32	0.01
HNSC & LUAD	5.20e-06	1.31e-8	0.035	9.8e-20	9.40e-18	0.035

NOTE: Both the known class labels ("True") and estimated labels ("Est") are used to calculate the cluster indices. Clustering errors are provided (defined at the beginning of this section).

Table 3. Application of the generalized SigClust-Soft and generalized MDS-based SigClust proposed in Section 2.5 on multi-cancer and a subset of breast cancer data.

		SigClust-Soft	SigClust-MDS
Multi cancer true $K = 3$	Decision	Reject H_0	Reject H_0
	Choice of K	2	3
Subset of breast cancer true $K = 4$	Decision	Reject H_0	Reject H_0
	Choice of K	5	3

Table 4. SigClust p-values for testing each single subgroup in the multi-cancer dataset.

	SigClust-Soft	SigClust-MDS
HNSC	2.87e-3	0.419
LUSC	0.351	0.954
LUAD	0.330	0.666

(LUAD). More information can be found in the Cancer Genome Atlas (TCGA) project (TCGA 2012). Each sample consists of 20531 genes estimated from RNA-seq data v2, which is available at *https://wiki.nci.nih.gov/display/TCGA/RNASeq+Version+2*. Following the same data preprocessing procedure as in Kimes et al. (2017), we use the log transformation of the original data and select a subset of 500 genes with the highest median absolute deviation (MAD) about the median. After preprocessing, the dataset consists of 300 samples and 500 genes.

Table 2 presents the testing results for pairwise subgroups using both estimated and given true labels, as well as the clustering errors. SigClust-soft and SigClust-MDS show high power (p-values ≈ 0) in all comparisons. The small p-values indicate that these clustering operations are statistically significant. The clustering errors are small for both methods and all three combinations of subgroups.

To implement the generalized MDS-based SigClust, we choose the set of test statistics CI_2, CI_3, CI_4, and CI_5 and set $r = 4$. Table 3 displays the performance of generalized SigClust methods. Both generalized SigClust-MDS and SigClust-Soft reject the null, while our SigClust-MDS successfully estimates the correct number of clusters as 3.

In addition, we apply our methods on every single subgroup, HNSC, LUSC, and LUAD. Within each group, we cluster the data into two parts to create artificial clusters to see whether our method can tell that the cluster operation within each class is not preferred. Table 4 shows that our SigClust-based MDS gives large p-values for all three cases, indicating each group should not be divided further.

5.2. Breast Cancer Gene Expression Dataset

We consider a gene expression dataset from 337 breast cancer samples which is categorized into five molecular subtypes: 97 LumA, 54 LumB, 91 basal-like, 47 normal breast-like, and 48 HER2-enriched samples. The dataset is available at *https://genome.unc.edu/pubsup/clow/*. We choose a subset of 1645 *intrinsic* genes identified in Prat et al. (2010).

Table 5 shows the p-values for 10 pairs of breast cancer subtypes. Both methods yield significant p-values for the first 8 pairs of comparison and insignificant p-values for the last pair. When testing the statistical significance of two breast cancer subtypes "LumA" and "LumB", our MDS-based SigClust gives insignificant p-values, suggesting that the two cancer subtypes are not significant clusters. This result is consistent with the fact that both "LumA" and "LumB" belong to the luminal cancer subtype. The luminal subtype has a big spectrum of samples but not necessarily contains two significant subgroups. According to Yersal and Barutca (2014), "LumA" and "LumB" have similar biological features with ER-responsive genes. Therefore, our MDS-based SigClust suggests it is not statistically significant to divide the luminal subtype into luminal A and luminal B.

For the pair of Her2 & LumB, SigClust-MDS with estimated labels gives a significant p-value while SigClust-Soft gives insignificant p-values. Figure 4(b) indicates that the two subgroups are separated in the first MDS direction although there is no significant gap between the two subgroups. Therefore, our MDS-based SigClust produces a more convincing testing result in this case.

To demonstrate the performance of the generalized MDS-based SigClust, we consider a subset consisting of four cancer subtypes "Basal," "Normal," "LumA," and "LumB" because of the overlaps between "Her2" and luminal groups, as shown in Figure 4(a). When applying the generalized MDS-based SigClust on the subset, our method estimates the number of clusters as 3, which is consistent with the pairwise testing result that "LumA" and "LumB" are not statistically different from each other, as shown in Table 3. The SigClust-soft estimates the number of clusters as 5, which is incorrect. The results for the entire dataset with all five subtypes are included in the Appendix. We can see from Table 3 of the Appendix that the clustering error is large (0.31) on the MDS space. Therefore, the evaluation result from the generalized MDS-based SigClust on the entire dataset is not reliable due to the poor clustering performance.

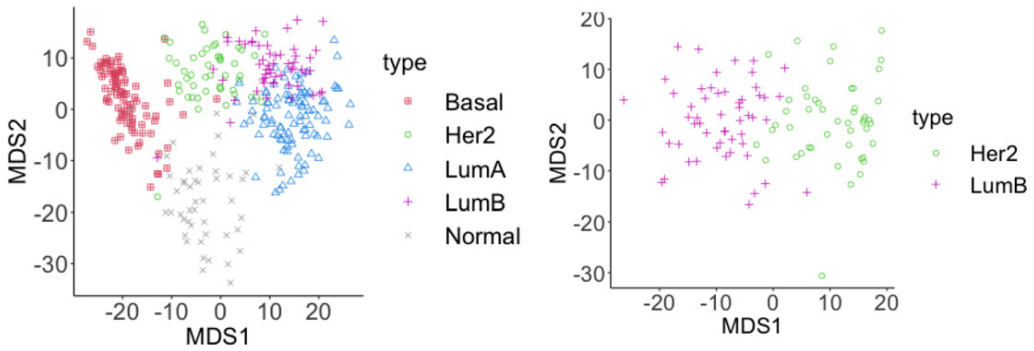
5.3. British Author Data

Our MDS-based SigClust is flexible because it can work with various distance functions. Here is an application where the Canberra distance handles count data. The Canberra distance d between vectors $\mathbf{p}$ and $\mathbf{q}$ in an n-dimensional real vector space is given as follows: $d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}$, where $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ and $\mathbf{q} = (q_1, q_2, \ldots, q_n)$ are vectors.

Table 5. SigClust *p*-values for each pair of subtypes for the breast cancer data.

	Soft(True)	Soft(Est)	Error(Soft)	MDS(True)	MDS(Est)	Error(MDS)
Basal & Normal	0.028	0.025	0.08	1.42e-05	2.63e-4	0.08
Basal & Her2	2.82e-3	1.66e-13	0.04	5.07e-12	4.86e-5	0.04
Basal & LumA	1.89e-8	2.70e-7	0.005	1.76e-39	6.83e-20	0.005
Basal & LumB	7.32e-6	1.483-4	0.01	1.68e-25	2.95e-11	0.01
Normal & Her2	0.034	0.018	0.07	9.28e-6	6.47e-5	0.07
Normal & LumA	9.18e-3	0.023	0.03	2.92e-7	4.3e-5	0.03
Normal & LumB	1.86e-3	3.19e-3	0.02	4.68e-14	5.03e-7	0.02
Her2 & LumA	4.0e-3	3.36e-12	0.02	1.49e-2	9.01e-6	0.03
Her2 & LumB	*0.220*	*0.282*	*0.069*	*0.084*	*0.022*	*0.078*
LumA & LumB	0.963	0.57	0.19	1	0.216	0.17

NOTE: Both known class labels ("True") and estimated labels ("Est") are used to calculate the CIs. Clustering errors are provided.



(a) Five types (b) Two types

Figure 4. The MDS projection scatterplots of the breast cancer data. Left: entire dataset with 5 cancer subtypes. Right: subset with 2 cancer types. True labels are used.

Table 6. SigClust *p*-values for each pair of subtypes for the British author data.

	Soft(True)	Soft(Est)	Error(Soft)	MDS(True)	MDS(Est)	Error(MDS)
Austen & London	5.72e-22	1.62e-5	0.09	3.23e-48	1.01e-6	0.09
Austen & Milton	7.98e-68	2.13e-59	0	4.02e-84	7.52e-26	0
Austen & Shakespeare	8.66e-56	8.95e-60	0.02	9.24e-71	5.59e-20	6e-3
London & Milton	*0.645*	*0.020*	*0.3*	*6.9e-37*	*9.41e-4*	*3e-3*
London & Shakespeare	9.85e-5	4.11e-7	0.06	6.73e-42	1.30e-14	0.06
Milton & Shakespeare	1.87e-35	3.58e-36	0.04	7.22e-34	8.87e-15	0

NOTE: Both known class labels ("True") and estimated labels ("Est") are used to calculate the cluster indices. Clustering errors are provided.

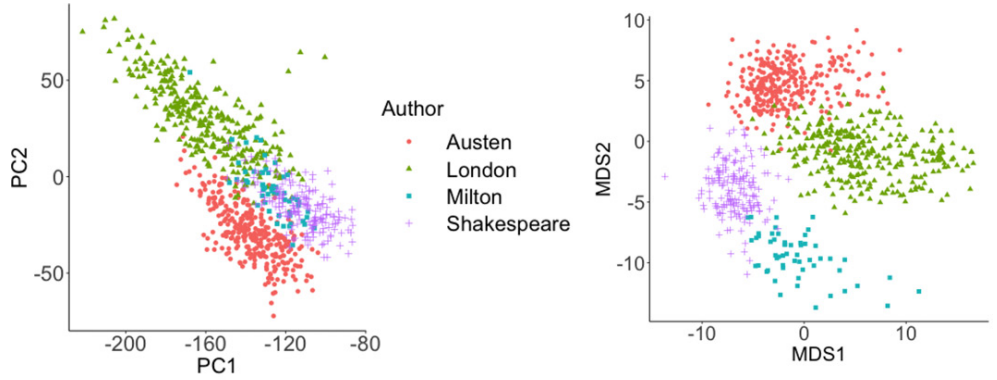
We implement the sample SigClust method, namely using the sample covariance matrix as an estimator of the population covariance matrix (denoted as SigClust-Sample) for comparison, because $d = 69 < n = 841$. This dataset consists of word counts from chapters written by four British authors: 317 chapters from Jane Austen, 296 from Jack London, 55 from John Milton, and 173 from William Shakespeare. The goal is to establish the statistical significance of clustering the dataset into subgroups according to the authors.

To demonstrate the usefulness of the Canberra distance, we visualize the data using the first two coordinates of the MDS matrix with the Euclidean and the Canberra distance. We use different shapes for different authors. As shown in Figure 5, the Canberra distance provides better separation among different authors than the Euclidean distance.

Table 5 shows the *p*-values for all pairs of authors based on MDS representations with the Canberra distance. Both methods yield significant *p*-values except the fourth pair, London & Milton. For this pair, SigClust-MDS gives a significant *p*-value while SigClust-Sample gives a large *p*-value (near 1) using the given labels. Figure 5 shows that the subgroups of London and Milton are mixed in the first two PC directions. Therefore, SigClust-Sample fails to identify subgroups under true labels. This application demonstrates that the Canberra distance helps to measure the difference among these subgroups and thus benefits the clustering evaluation task.

5.4. Applications in Natural Language Processing

As discussed before, our MDS-based SigClust works even when the original data are unavailable, as long as the dissimilarity matrix is provided. In this analysis, we apply the MDS-based SigClust on canonical natural language datasets available from Nakamura (2006). For natural language terms, the data points do not have geometric coordinates. Thus, these datasets are in the form of distance matrices where the Google distance captures the pairwise distance. To test the significance of clusters, we apply the base version of our MDS-based SigClust with r=2 on each pair of subgroups. Six datasets **people5**, **alt-ds**, **math-med-fin**, **finance-cs-j**, **phil-avi-d** and **math-cuisine** are analyzed. Detailed descriptions of datasets are given in the Appendix.



(a) PCA projection scatterplot

(b) MDS projection with Canberra distance

Figure 5. PCA and MDS projection scatterplot view of the British author data. True labels are used.

Table 7. MDS-based SigClust test results for the dataset people5.

	SigClust-MDS True	SigClust-MDS Est	clustering error
Classical composers vs artists	3.77e-12	6.43e-13	0
Classical composers vs authors	4.54e-10	3.52e-10	0.04
Classical composers vs math	8.64e-12	2.91e-11	0
Classical composers vs pop music	1.87e-08	5.83e-08	0.02
Artists vs authors	2.21e-10	4.62e-09	0
Artists vs mathematicians	2.87e-09	2.11e-09	0.02
Artists vs pop musicians	1.04e-12	5.84e-14	0
Authors vs mathematicians	2.29e-05	7.03e-08	0.06
Authors vs pop musicians	0.071	0.02	0.04
Mathematicians vs pop musicians	2.37e-06	2.43e-06	0.04

Table 7 displays the clustering and MDS-based SigClust test results for the dataset **people5** and the results for the other datasets are in Table 5 of the Appendix. The "SigClust-MDS True" and "SigClust-MDS Est" columns show the testing results where cluster indices are calculated using true and estimated labels, respectively. The "clustering error" and "misclassified nodes" columns display the error rates and the numbers of misclassified nodes under the 2-means algorithm.

As shown in Table 7 (and Table 5 of the Appendix), the clustering error is very small in each case. Therefore, the MDS matrix gives a good two-dimensional representation of the original distance matrix and preserves the distance information. The testing results for datasets **alt-ds, math-med-fin, phil-avi** and **math-cuisine** are significant (< 0.05) using both true labels and estimated labels. The only two insignificant tests are the "author vs. pop musicians" comparison in **people5** and the one in **finance-cs-j** using true labels. Both cases have p-values slightly larger than 0.05. In summary, our MDS-based SigClust suggests that those clusters are mostly significantly different from each other.

6. Discussion

In this article, we propose a new MDS-based SigClust method for testing the statistical significance of clustering. Our method combines the original SigClust method and multidimensional scaling. The most challenging part of the original SigClust is the estimation of the high-dimensional covariance matrix. Furthermore, one may only have the pairwise dissimilarity matrix between samples available without the original data in various applications. Our new method can tackle these challenges effectively. We can obtain low-dimensional representations of the original data through MDS using only the dissimilarity matrix. Through extensive simulation studies, we show that the MDS matrix can preserve existing cluster information under null and alternative hypotheses. Our method can control Type I error under the null and is powerful under the alternative hypothesis. As an extension of the original cluster index, the combined cluster index successfully captures separation signals from all possible directions. The extension makes our MDS-based SigClust more broadly applicable. Moreover, we propose a generalized MDS-based SigClust that can identify the number of clusters when there is more than one.

There are several open directions for future research. One interesting direction is to further develop theoretical results under more general null distributions in contrast to the usual Gaussian cluster assumption. Another important area of research is to establish consistency of the estimated number of clusters using generalized MDS-based SigClust.

Supplementary Materials

Appendix: Proofs of all theoretical results for Section 3 and additional numerical results for Sections 4 and 5. (Appendix.pdf, pdf file)

Code and data: Example code and code for each section in the manuscript and the Appendix. Please read the file README contained in the zip file for more details. (code_and_data.zip, zip file)

Acknowledgments

The authors are indebted to the editor, the associate editor, and two reviewers, whose helpful suggestions led to a much improved presentation.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

The authors were supported in part by NSF grants DMS-2113662 (Bhamidi and Shen), DMS-1613072 (Bhamidi), DMS-1606839 (Bhamidi), DMS-2134107 (Bhamidi), DMS-2100729 (Liu), SES-2217440 (Liu); ARO grant W911NF-17-1-0010 (Bhamidi); and NIH grant R01-GM126550 (Liu and Shen).

ORCID

Yufeng Liu https://orcid.org/0000-0002-1686-0545

References

Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2020), "Entrywise Eigenvector Analysis of Random Matrices with Low Expected Rank," *Annals of Statistics*, 48, 1452–1474. [222]

Agrawal, N., Akbani, R., Aksoy, B. A., Ally, A., Arachchi, H., Asa, S. L., Auman, J. T., Balasundaram, M., Balu, S., Baylin, S. B. et al. (2014), "Integrated Genomic Characterization of Papillary Thyroid Carcinoma," *Cell*, 159, 676–690. [220]

Ben-Hur, A., Horn, D., Siegelmann, H. T., and Vapnik, V. (2001), "Support Vector Clustering," *Journal of Machine Learning Research*, 2, 125–137. [219]

Borg, I., and Groenen, P. J. F. (2005), *Modern Multidimensional Scaling Theory and Applications*, New York: Springer. [220,221]

Chakravarti, P., Balakrishnan, S., and Wasserman, L. (2019), "Gaussian Mixture Clustering Using Relative Tests of Fit," arXiv preprint arXiv:1910.02566. [219,220,222,224,225]

Fraley, C., and Raftery, A. E. (2002), "Model-based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631. [220]

Fred, A. L., and Jain, A. K. (2005), "Combining Multiple Clusterings Using Evidence Accumulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 835–850. [219]

Holm, S. (1979), "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 6, 65–70. [223]

Huang, H., Liu, Y., Yuan, M., and Marron, J. (2015), "Statistical Significance of Clustering Using Soft Thresholding," *Journal of Computational and Graphical Statistics*, 24, 975–993. [220,224]

Kimes, P. K., Liu, Y., Neil Hayes, D., and Marron, J. S. (2017), "Statistical Significance for Hierarchical Clustering," *Biometrics*, 73, 811–821. [220,227]

Little, A., Xie, Y., and Sun, Q. (2022), "An Analysis of Classical Multidimensional Scaling with Applications to Clustering," *Information and Inference: A Journal of the IMA*, 12, 72–112. [223]

Liu, Y., Hayes, D. N., Nobel, A., and Marron, J. S. (2008), "Statistical Significance of Clustering for High-Dimension, Low–Sample Size Data," *Journal of the American Statistical Association*, 103, 1281–1293. [219,220,221]

Löffler, M., Zhang, A. Y., and Zhou, H. H. (2021), "Optimality of Spectral Clustering in the Gaussian Mixture Model," *Annals of Statistics*, 49, 2506–2530. [222]

MacQueen, J. et al. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Oakland, CA, USA (Vol. 1), pp. 281–297. [219]

Maitra, R., Melnykov, V., and Lahiri, S. N. (2012), "Bootstrapping for Significance of Compact Clusters in Multidimensional Datasets," *Journal of the American Statistical Association*, 107, 378–392. [219]

McLachlan, G., and Peel, D. (2000), *Finite Mixture Models*, Wiley Series in Probability and Statistics, pp. 420–427, New York: Wiley. [220]

McShane, L. M., Radmacher, M. D., Freidlin, B., Yu, R., Li, M.-C., and Simon, R. (2002), "Methods for Assessing Reproducibility of Clustering Patterns Observed in Analyses of Microarray Data," *Bioinformatics*, 18, 1462–1469. [219,224]

Nakamura, A. (2006), "Laboratory for Algorithmics: Datasets," available at *https://www-alg.ist.hokudai.ac.jp/datasets.html*. [220,228]

Poland, J., and Zeugmann, T. (2006), "Clustering Pairwise Distances with Missing Data: Maximum Cuts Versus Normalized Cuts," in *International Conference on Discovery Science*, pp. 197–208, Springer. [220]

Pollard, D. (1982), "A Central Limit Theorem for *k*-means Clustering," *Annals of Probability*, 10, 919–926. [223]

Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., He, X., and Perou, C. M. (2010), "Phenotypic and Molecular Characterization of the Claudin-Low Intrinsic Subtype of Breast Cancer," *Breast Cancer Research*, 12, R68. [227]

TCGA (2012), "Comprehensive Genomic Characterization of Squamous Cell Lung Cancers," *Nature*, 489, 519–525. [220,227]

Von Luxburg, U. (2007), "A Tutorial on Spectral Clustering," *Statistics and Computing*, 17, 395–416. [219]

Xu, D., and Tian, Y. (2015), "A Comprehensive Survey of Clustering Algorithms," *Annals of Data Science*, 2, 165–193. [219]

Yersal, O., and Barutca, S. (2014), "Biological Subtypes of Breast Cancer: Prognostic and Therapeutic Implications," *World Journal of Clinical Oncology*, 5, 412–24. [227]