Clustering analysis of multivariate data

Baragilly, Mohammed; Gabr, Hend; Willis, Brian H

DOI:
10.1155/2023/8849404

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Baragilly, M, Gabr, H & Willis, BH 2023, 'Clustering analysis of multivariate data: a weighted spatial ranks-based approach', *Journal of Probability and Statistics*, vol. 2023, 8849404. https://doi.org/10.1155/2023/8849404

Hindawi
Journal of Probability and Statistics
Volume 2023, Article ID 8849404, 15 pages
https://doi.org/10.1155/2023/8849404

Research Article

Clustering Analysis of Multivariate Data: A Weighted Spatial Ranks-Based Approach

Mohammed H. Baragilly,[1,2] **Hend Gabr**,[3,4] **and Brian H. Willis**[2]

[1]*Department of Mathematics, Insurance and Applied Statistics, Helwan University, Helwan, Egypt*
[2]*Institute of Applied Health Research, University of Birmingham, Birmingham, UK*
[3]*Department of Mathematics, Insurance, and Statistics, Faculty of Commerce, Menoufia University, Shibin El Kom, Egypt*
[4]*Centre for Women's Mental Health, University of Manchester, Manchester, UK*

Correspondence should be addressed to Brian H. Willis; b.h.willis@bham.ac.uk

Received 15 December 2022; Revised 27 March 2023; Accepted 5 June 2023; Published 30 September 2023

Academic Editor: Hyungjun Cho

Determining the right number of clusters without any prior information about their numbers is a core problem in cluster analysis. In this paper, we propose a nonparametric clustering method based on different weighted spatial rank (WSR) functions. The main idea behind WSR is to define a dissimilarity measure locally based on a localized version of multivariate ranks. We consider a nonparametric Gaussian kernel weights function. We compare the performance of the method with other standard techniques and assess its misclassification rate. The method is completely data-driven, robust against distributional assumptions, and accurate for the purpose of intuitive visualization and can be used both to determine the number of clusters and assign each observation to its cluster.

1. Introduction

In recent years, there has been a significant advancement in clustering methods in the field of data science and machine learning. Density-based spatial clustering of applications with noise (DBSCAN) has become one of the popular nonparametric clustering algorithms that group together points that are close to each other and are surrounded by low-density areas [1]. Also, hierarchical clustering [2] has been widely used to build a hierarchy of clusters by either merging smaller clusters into larger ones or splitting larger clusters into smaller ones. The K-means clustering [3] remains one of the most popular partitioning methods, where the data are partitioned into K distinct, nonoverlapping clusters.

In addition, spectral clustering [4] has gained popularity due to its ability to use the eigenvectors of the similarity matrix to partition the data into clusters. Affinity propagation [5] assigns each data point to an exemplar, which is a representative point of a cluster, and iteratively updates the exemplars until convergence. Fuzzy clustering [6] assigns each data point a probability of belonging to each cluster, rather than assigning it to a single cluster. The K-medoids algorithm introduced by [7] uses actual data points as representatives or medoids for each cluster, and medoids selected from the dataset are typically the most centrally located points within a cluster. This makes K-medoids more robust to outliers and noise compared to K-means. Another recent clustering method is the distance density clustering (DDC) introduced by [8]. It is a distance density clustering method that is a medoid-based clustering with time series data density consideration which provides clustering results in a hierarchy fashion.

The clustering by the fast search and find of density peaks (densityClust) method [9] is a density-based clustering method that aims to identify clusters based on the local density of data points and the distances between them, which makes it suitable for datasets with irregularly shaped clusters and varying densities. Furthermore, neighborhood grid clustering (NGC) introduced by [10] is a density-based

clustering algorithm that aims to identify clusters based on the local density of data points and their spatial relationships. It utilizes a grid-based approach to partition the data space into cells and then assigns data points to their corresponding cells. The algorithm then determines the cluster assignments by considering the density and spatial proximity of the points within each cell.

These methods have their own strengths and weaknesses and are suitable for different types of data and applications. The choice of the appropriate clustering method depends on various factors such as the size and nature of the data, the desired number of clusters, and the research question being addressed.

Further advancements have been made in the use of spatial ranks to analyse multivariate data, with the development of nonparametric methods being particularly notable [11–13]. They have a number of attractive features including being distribution-free and easy to compute. Furthermore, the traditional spatial ranks function gives information about how central each observation is and its direction in relation to the centre. However, they do not capture the distances between each pair of observations, which is important for cluster analysis.

Recently, Baragilly and Chakraborty [11] utilized spatial ranks as a clustering tool, using a forward search methodology based on nonparametric multivariate spatial rank functions to determine the number of clusters in the data. Their method does not depend on the choice of the initial subsample and has been shown to perform well in different mixture distributions. This work has been extended to clustering functional datasets in medical applications [14].

This paper proposes a novel approach to utilizing spatial ranks for clustering, using a nonparametric weighted spatial ranks function that takes into account the distances between each pair of observations as weights and defines a dissimilarity measure based on spatial ranks. By measuring the distances between each pair of observations instead of their central tendency, it becomes easier to segment a given set of data into a specific number of clusters.

The main idea behind the weighted spatial ranks (WSRs) is to define a dissimilarity measure based on a localized version of spatial ranks, such that the weighted ranks can be used as a classifier and a confirmatory tool to determine the number of clusters and assign each observation to its cluster. Proper selection of a weight function can lead to better identification of clusters, and kernel weights are a popular choice in pattern analysis, classification, cluster analysis, machine learning, and support vector machines.

The paper also demonstrates how weighted spatial ranks may be used for the purpose of visualizing the clusters, so that the number of clusters may be determined using weighted rank contours for a low-dimensional input space after dimension reduction.

Section 2 of the paper introduces the weighted spatial rank function and evaluates its use for different parametric and nonparametric weight functions. Section 3 demonstrates the weighted rank-based clustering algorithm and proposes a confirmatory classifier based on weighted ranks that can be used to assign observations to the most appropriate cluster for two-dimensional data. Section 4 demonstrates the weighted rank-based clustering algorithm to higher dimensional data, and Section 5 provides numerical examples based on both simulated and real datasets to examine the performance of the proposed algorithm. The algorithm is compared with other clustering methods in Section 6, and concluding remarks are presented in Section 7.

2. Weighted Spatial Rank Functions

In this section, we propose two different weighted spatial rank functions. Suppose that $\mathbf{X} \in \mathbb{R}^d$ has a d-dimensional distribution of F, which we assume to be continuous hereon, then the unweighted spatial rank function of the point $\mathbf{X} \in \mathbb{R}^d$ with respect to F can be defined as

$$\begin{aligned}\mathbf{SR}_{F_n}(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^{n}\operatorname{Sign}(\mathbf{x}-\mathbf{X}_i)\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{x}-\mathbf{X}_i}{\|\mathbf{x}-\mathbf{X}_i\|}.\end{aligned} \tag{1}$$

The first weighted spatial rank function of the point $\mathbf{X} \in \mathbb{R}^d$ with respect to F is a vector function and can be defined as

$$\begin{aligned}\mathbf{WSR}^{(1)}_{F_n}(\mathbf{x}) &= \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i \operatorname{sign}(\mathbf{x}-\mathbf{X}_i)\\ &= \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i \frac{\mathbf{x}-\mathbf{X}_i}{\|\mathbf{x}-\mathbf{X}_i\|}.\end{aligned} \tag{2}$$

The second weighted spatial rank function can be defined as

$$\mathbf{WSR}^{(2)}_{F_n}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} w_i \operatorname{sign}(\mathbf{x}-\mathbf{X}_i) = \frac{1}{n}\sum_{i=1}^{n} w_i \frac{\mathbf{x}-\mathbf{X}_i}{\|\mathbf{x}-\mathbf{X}_i\|}, \tag{3}$$

where their L_2 norms are

$$\mathbf{WSRN}^{(1)}_{F_n}(\mathbf{x}) = \left\|\mathrm{WSR}^{(1)}_{F_n}(\mathbf{x})\right\|, \tag{4}$$

$$\mathbf{WSRN}^{(2)}_{F_n}(\mathbf{x}) = \left\|\mathrm{WSR}^{(2)}_{F_n}(\mathbf{x})\right\|. \tag{5}$$

Note that, the difference between (2) and (3) lies with the scale (denominator) which in (2) is the sum of the weights of w_i where $w_i = w_i(\mathbf{x})$ and is therefore data dependent, whereas in (3) it depends on n which is data independent.

Kernel weight functions are often used in nonparametric estimation and, as already indicated, in a range of classification and pattern recognition problems. Here, we consider the Gaussian kernel weight, which is one of the commonly used nonparametric kernel weight functions and it is defined as follows:

$$w_i = e^{-\|\mathbf{x}-\mathbf{X}_i\|^2/2}, \tag{6}$$

where $\|\mathbf{x}\|$ is the Euclidean norm such that $\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_d^2}$ is the direction of the d-dimensional vector $\mathbf{x}$ (for comprehensive details on kernel weights, see Souza [15])

3. Weighted Spatial Ranks Clustering Algorithm

We now introduce the weighted spatial ranks clustering algorithm starting with the bivariate case ($d = 2$) before considering the higher dimensional case $d > 2$.

3.1. Weighted Spatial Ranks Clustering Algorithm for a Bivariate Case ($d = 2$)

(1) Let $\mathbf{X}_1, \mathbf{X}_2, \ldots \mathbf{X}_n \in \mathbb{R}^2$ be a random sample with two variables x_1 and x_2 and let $\mathbf{S}_\mathbf{x}$ be the Cartesian product of two equally spaced sets, $S_{x_1} = \{\min(x_1), \ldots, \max(x_1)\}$ and $S_{x_2} = \{\min(x_2), \cdots, \max(x_2)\}$ so that each $\mathbf{s} \in \mathbf{S}_\mathbf{x}$ is a two-dimensional vector.

(2) For each $\mathbf{s} \in \mathbf{S}_x$, we calculate $\mathrm{WSRN}(\mathbf{s})$ with respect to $\mathbf{X}_i$as

$$\begin{aligned}\mathrm{WSRN}(\mathbf{s}) &= \|\mathbf{WSR}(\mathbf{s})\| \\ &= \left\| \frac{1}{n} \sum_{i=1}^{n} w_i \frac{\mathbf{s} - \mathbf{X}_i}{\|\mathbf{s} - \mathbf{X}_i\|} \right\|,\end{aligned} \tag{7}$$

where $i = 1, \ldots n$ and $w_i = e^{-\|\mathbf{s} - \mathbf{X}_i\|^2/2}$.

(3) Plot the weighted functional spatial ranks contour based on $\mathrm{WSRN}(\mathbf{s})$ and $s \in \mathbf{S}_\mathbf{x}$, and then determine the number of clusters K from the contour lines.

(4) Based on the contour lines, specify the assigned observations in each cluster. You can use a lower contour level for better visualization.

(5) Use the weighted spatial rank classifier's rule defined in Section 3.2 to confirm the assignment of each observation and allocate the unassigned observations to the proper cluster.

3.2. Weighted Spatial Ranks Classifier. Suppose that we have k groups, with distributions $F_1, F_2, \ldots, F_k$, then based on the second WSR function in (3), we can assign d-dimensional observation vector $\mathbf{x}$ to the i-th group if

$$\mathrm{WSRN}_{F_i}^{(2)}(\mathbf{x}) = \max_{1 \le j \le k} \mathrm{WSRN}_{F_j}^{(2)}(\mathbf{x}), \tag{8}$$

where $i \neq j, 1 \le i \le k$. Note, if we had used $\mathrm{WSR}_{F_i}^{(1)}$, then since $\mathrm{WSR}_{F_i}^{(1)}$ increases outward from the spatial median, we have

$$\mathrm{WSRN}_{F_i}^{(1)}(\mathbf{x}) = \min_{1 \le j \le k} \mathrm{WSRN}_{F_j}^{(1)}(\mathbf{x}). \tag{9}$$

So, after determining the number of clusters using the WSR contour, (8) may be used to assign each observation to the most suitable cluster.

3.3. Confirmatory Analysis Based on the Weighted Spatial Ranks Clustering Algorithm. In order to assess the weighted spatial rank functions' performance, we compare them with other standard methods such as Euclidean distances, Mahalanobis distances, spatial ranks, and spatial depth. A simulation study is used to assess the performance of the proposed weighted ranks-based clustering in defining the group structure in multivariate data. The simulated data are sampled from a bivariate normal mixture distribution that is assumed to cluster into two groups, where the mixture proportion $p = 0.3$ and sample size $n = 1000$ such that $\mathbf{X}_1$, $\mathbf{X}_2, \ldots, \mathbf{X}_n \in \mathbb{R}^d$ is a random sample from

$$p.N_2(\boldsymbol{\mu}_1, \Sigma) + (1 - p).N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}), \tag{10}$$

where $\boldsymbol{\mu}_1 = (0, 0)^T$, $\boldsymbol{\mu}_2 = (5, 5)^T$, and $\boldsymbol{\Sigma} = \mathbf{I}$.

For all the contour plots that follow, they are derived from a random sample of 1000 observations simulated from the bivariate mixture normal distribution in (10). Figure 1 shows the contour plots of the Euclidean distances, Mahalanobis distances, spatial ranks, and the spatial depth. The figure clearly shows that the contours produced have failed to map the shape of the two clusters' structure in the bivariate mixture distribution.

In Figure 2, $\mathrm{WSRN}_{F_n}^{(1)}$ (defined in (4)) and $\mathrm{WSRN}_{F_n}^{(2)}$ (defined in (5)) are used to derive the contour plots based on the nonparametric Gaussian kernel weights function defined in (6) (see section 4). In general, Figure 2 reveals that compared with Figure 1, the contours produced from both $\mathrm{WSRN}_{F_n}^{(1)}$ and $\mathrm{WSRN}_{F_n}^{(2)}$ based on the Gaussian kernel function capture more of the structure of the simulated data. However, the contour based on $\mathrm{WSRN}_{F_n}^{(1)}$ failed to detect some of the observations that are not close to either of the clusters. These undetected observations are indicated by some lines between the two clusters, suggesting the potential presence of a third cluster. In contrast, it is clear that the cluster structure of the data is better defined by using the $\mathrm{WSRN}_{F_n}^{(2)}$ as compared to that from the $\mathrm{WSRN}_{F_n}^{(1)}$ in Figure 2(a).

This is because $\mathrm{WSRN}_{F_n}^{(1)}$ is a constant and the values increase outward from the centre or spatial median of the cluster. Thus, unassigned points may be assigned to that cluster which results in the lowest weighted rank for the point. In contrast, the contour derived from the second normed weighted spatial rank function, $\mathrm{WSRN}_{F_n}^{(2)}$, defined in (5), is based on the values of $\mathrm{WSRN}_{F_n}^{(2)}$ that decrease outward from the spatial median. Thus, the larger the weighted spatial rank value for an individual point, the closer it is to the centre of the cluster.

In summary, $\mathrm{WSRN}_{F_n}^{(2)}$ is more successful than $\mathrm{WSRN}_{F_n}^{(1)}$ at capturing the cluster structure of the simulated data. Overall, the most accurate contour plot is synthesized when $\mathrm{WSRN}_{F_n}^{(2)}$ uses Gaussian kernel weights.

4. Weighted Spatial Ranks Based Clustering Algorithm for Higher Dimensions ($d > 2$)

For real-world datasets, we often have to analyse complex multidimensional data and this makes data visualization and computation more complicated. In such cases, a dimension

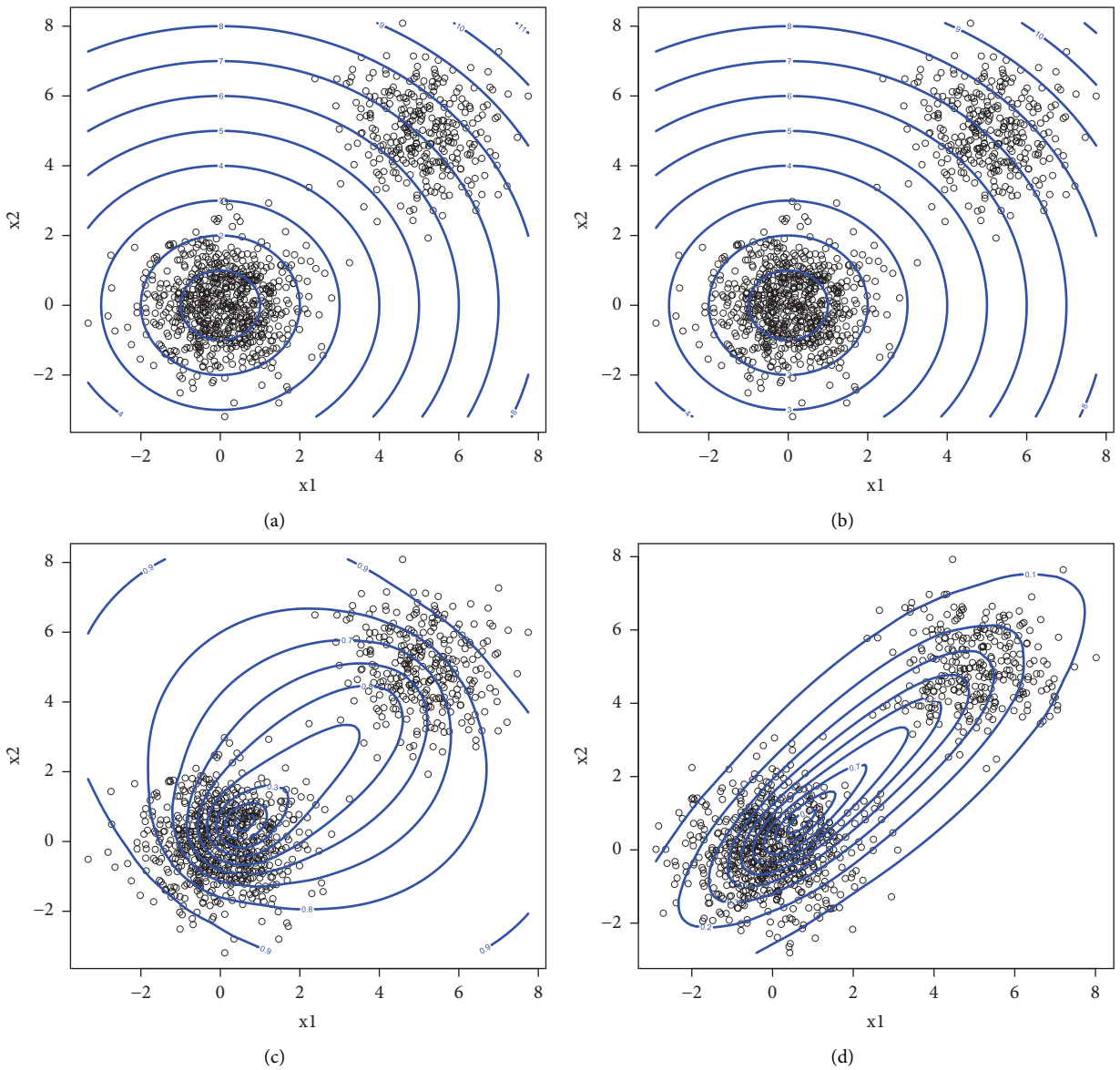


Figure 1: Simulated data example: contour plots of (a) Euclidean distances, (b) Mahalanobis distances, (c) spatial ranks, and (d) spatial depth based on 1000 random observations from bivariate mixture normal distribution with the two groups.

reduction strategy may be employed and here we use the principal component analysis (PCA) to reduce the dimensionality of the data to two dimensions in order to derive a contour plot (see [16]).

The main idea of using PCA is to find a lower-dimensional subspace that captures most of the variance in the data. Specifically, it involves finding the orthogonal rotation of the axes that maximises the variance. For a d-dimensional random variable $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ with covariance matrix Σ, let $C_j = a_j^T \mathbf{X}^{\mathbf{T}}$, where a_j is a d-dimensional vector of constants. Since the variance of $C_j = a_j^T \sum a_j$, to find the principal component, C_j requires finding

$$\arg\max_{a_j} \left(a_j^T \sum a_j\right) \text{ subject to } a_j^T a_j = 1. \tag{11}$$

It is necessary to constrain a_j to have a unit length to ensure finite values. This may be solved by using the method of Lagrange multipliers, so that for Lagrange multiplier, λ_j, this reduces to solving the eigen equation.

$$\left(\sum -\lambda_j I_d\right) a_j = 0. \tag{12}$$

It also means that for the component which yields the largest eigenvalue, λ_j has the largest variance and a_j is the corresponding eigenvector. It is also straightforward to show that the a_j for $j = 1, \ldots, d$ are orthogonal. Thus, for the $d \times d$ matrix of eigenvectors, $\mathbf{A} = [a_1, \ldots, a_d]$ and the matrix of principal component scores $\mathbf{C}$ is given by $\mathbf{C} = \mathbf{XA}$. This represents a rigid rotation of the $\mathbf{X}$ axes to an orientation of maximum variance. Thus, the first principal component C_1 has the largest variance for $\mathbf{X}$, and the second component C_2

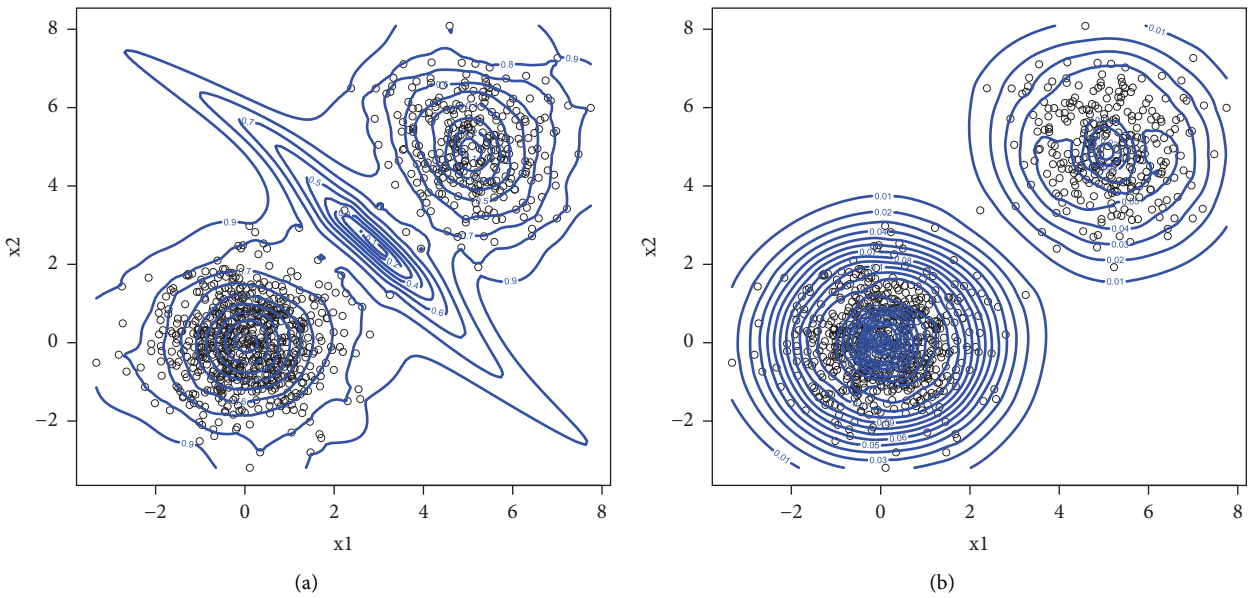


FIGURE 2: Simulated data example: contour plots of (a) the normed weighted spatial rank function $\mathrm{WSRN}_{F_n}^{(1)}$ and (b) the normed weighted spatial rank function $\mathrm{WSRN}_{F_n}^{(2)}$, using Gaussian kernel weights based on 1000 random observations from bivariate mixture normal distribution with the two groups.

has the second largest variance and so on (for more details on PCA, see [17]).

4.1. Weighted Spatial Ranks Clustering Algorithm When $d > 2$

(1) Let $\mathbf{X}_1, \mathbf{X}_2, \ldots \mathbf{X}_n \in \mathbb{R}^d$ be a d-dimensional random sample, then use the PCA to get the first two components $\mathbf{C}_1$ and $\mathbf{C}_2$ of that sample and construct the matrix $\mathbf{C}$, which is a matrix consisting of the two components $\mathbf{C}_1$ and $\mathbf{C}_2$.

(2) Consider $\mathbf{C}_1$ and $\mathbf{C}_2$ as the new variables and perform the steps of the weighted spatial ranks clustering algorithm when $d = 2$.

5. Numerical Examples

In this section, we apply the weighted spatial ranks-based clustering algorithm to two simulated datasets and three real datasets.

5.1. Simulated Data Examples. In the first simulated data example, we consider a mixture of three quadvariate normal distributions, with mixing proportions, $p_1 = 0.3$ and $p_2 = 0.4$, and sample size $n = 100$, such that $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \in \mathbb{R}^d$ is a random sample from 4-dimensional mixture normal distribution.

$$p_1 N_4(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + p_2 N_4(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) + (1 - p_1 - p_2) N_4(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}), \tag{13}$$

where $\boldsymbol{\mu}_1 = (4, 4, 4, 4)^T, \boldsymbol{\mu}_2 = (-4,4, -4,4)^T$, $\boldsymbol{\mu}_3 = (-4, -4, -4, -4)^T$, and $\boldsymbol{\Sigma} = \mathbf{I}$.

Figure 3(a) shows the scatterplot matrix of the principal components and reveals a mixed picture. It is clear from the component 1 versus component 2 panels that there are 3 clusters. In contrast, the component 2 versus component 3 and the component 2 versus component 4 suggest that there are only 2 clusters. Figure 3(b) gives the proportion of the total variance explained by each component, i.e., 97% of the total variance is explained by the first two components. Figures 3(c) and 3(d) demonstrate that the weighted spatial ranks contour accurately fits the shape of the three clusters without any misclassification. Finally, in Figures 3(e) and 3(f), the confirmatory plots based on the weighted ranks classifier for the first two components show that the observations have been correctly assigned to the three simulated clusters.

In the second example, we simulate a random sample of size $n = 100$ from a mixture of four 6-dimensional normal distributions, with equal proportions of weights $p = 0.25$, i.e.,

$$pN_6(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + pN_6(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) + pN_6(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}) + pN_6(\boldsymbol{\mu}_4, \boldsymbol{\Sigma}), \tag{14}$$

with $\boldsymbol{\mu}_1 = (4, 4, 4, 4, 4, 4)^T, \boldsymbol{\mu}_2 = (-4, 4, -4, 4, -4, 4)^T$, $\boldsymbol{\mu}_3 = (-4, -4, -4, -4, -4, -4)^T$, $\boldsymbol{\mu}_4 = (4, -4, 4, -4, 4, -4)^T$, and $\boldsymbol{\Sigma} = \mathbf{I}$.

From the scatter plot matrix in Figure 4(a), we can see that whilst there are four clear clusters in the component 1 versus component panels, the number of clusters is less clear in the other panels. However, it is clear from Figure 4(b) that the first two components explain the majority (98%) of the variance.

The weighted spatial ranks contour plots shown in Figures 4(c) and 4(d) clearly reveal the shape of the four clusters, where in the latter, a lower contour level = 0.001 has

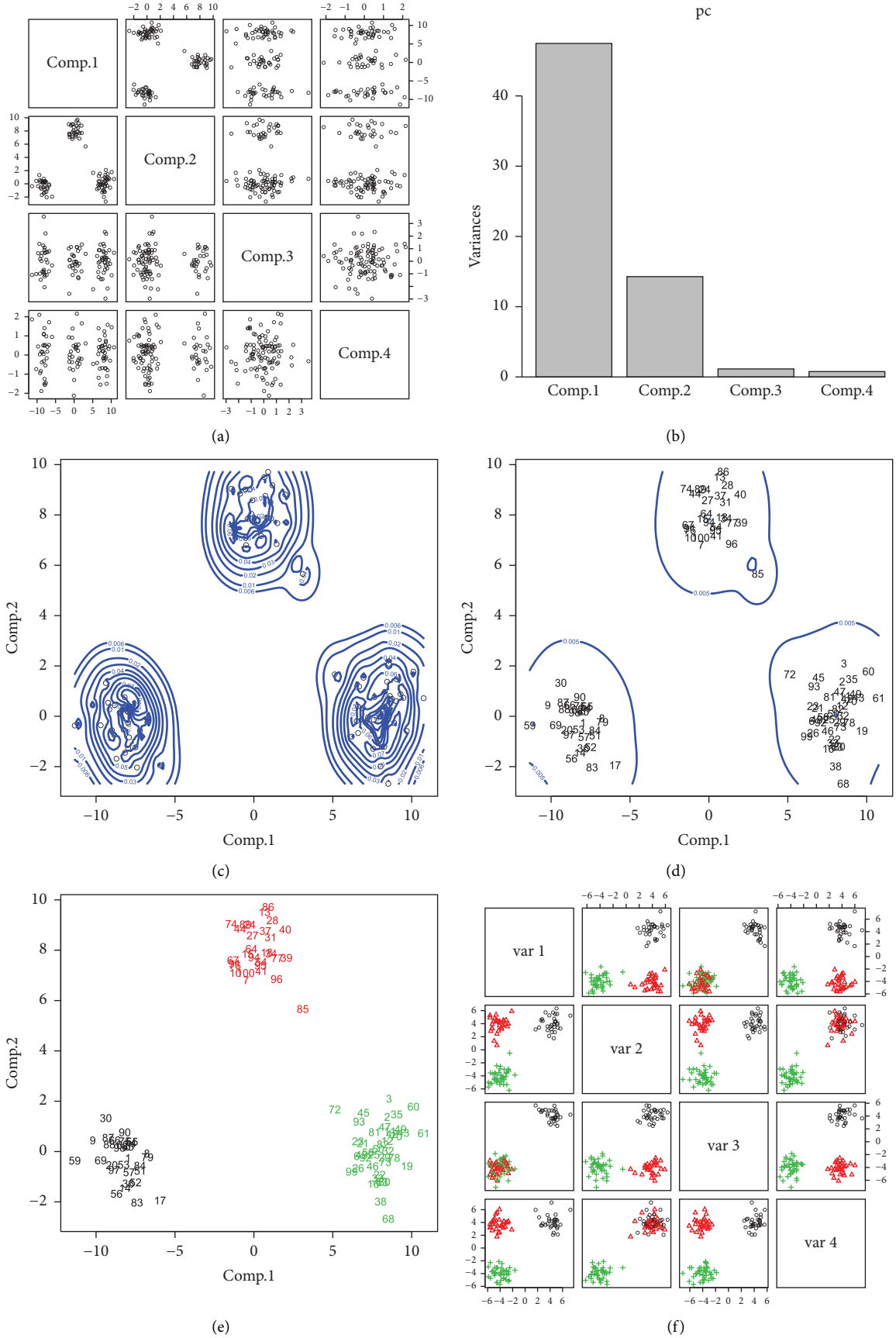


Figure 3: Simulated data 1: (a) scatterplot matrix of the PCA components, (b) the total variance explained by each component, (c) the weighted spatial ranks contour, (d) the contour at level 0.005 and the confirmatory plots based on weighted ranks classifier for (e) the first 2 components and (f) the original data.

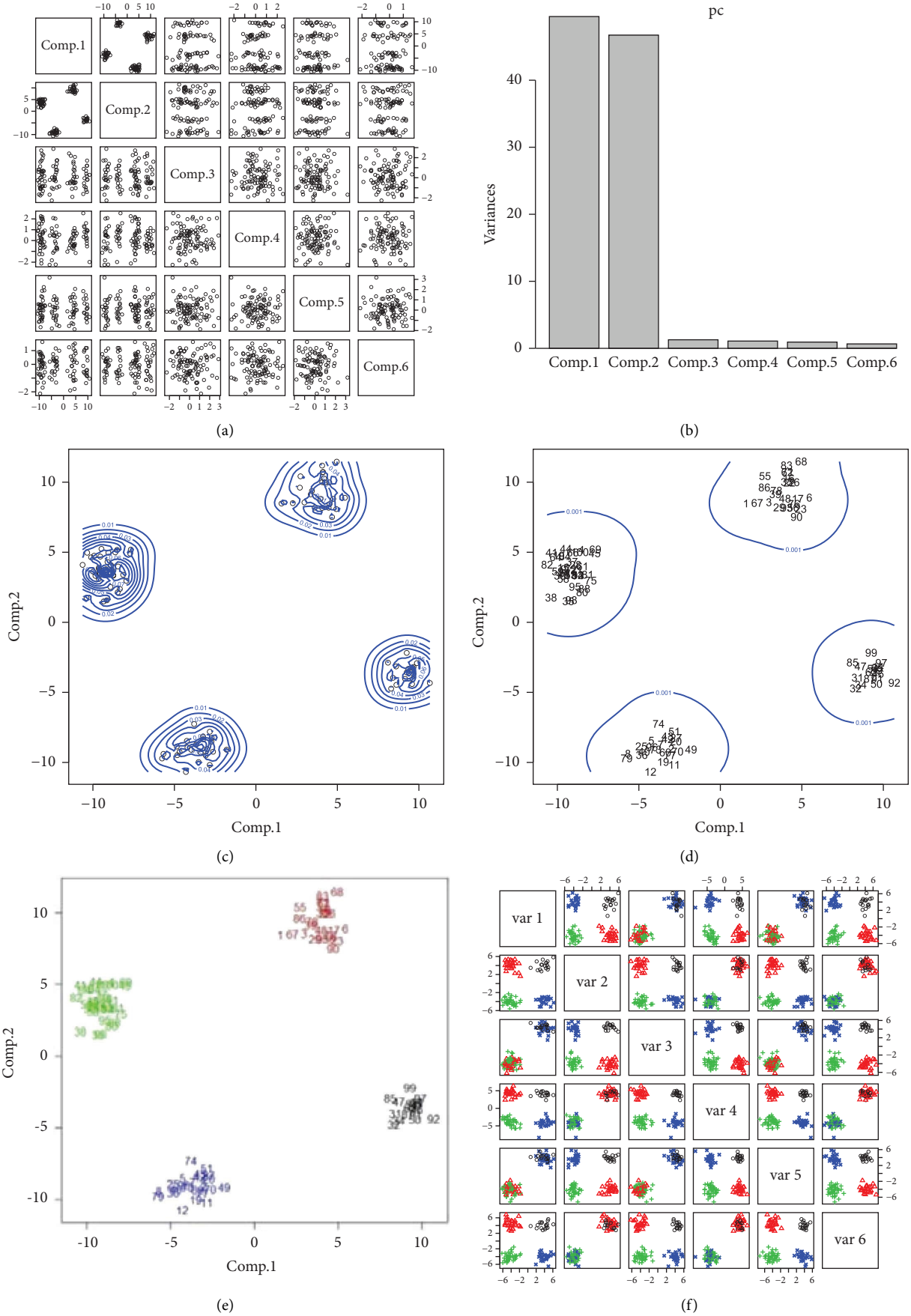


Figure 4: Simulated data 2: (a) scatterplot matrix of the PCA components, (b) the total variance explained by each component, (c) the weighted spatial ranks contour, (d) the contour at level 0.001 and the confirmatory plots based on weighted ranks classifier for (e) the first 2 components and (f) the original data.

been used. Finally, the confirmatory plots based on the weighted spatial ranks classifier for the first two components and the original data shown in Figures 4(e) and 4(f) demonstrate the correct assignment of the simulated observations to the right clusters.

5.2. Real Datasets' Examples. In this subsection, the algorithm is applied to three real datasets: the iris data [18], financial data [19], and old faithful geyser data [20, 21]. The iris dataset consists of three different types of irises (Setosa, Versicolour, and Virginica). However, most of the clustering techniques consider there to be two groups, since iris Virginica and iris Versicolour are not separable without the species information that Fisher used. As we can see from Figure 5(c), the weighted ranks contour based on the first two components, which explain 97.8% of the total variances, indicates two clusters. The confirmatory plots in Figures 5(e) and 5(f) assign all of the observations to two groups.

The second real dataset is the financial data [19], which contains measurements of the three variables monitoring the performance of 103 investment funds operating in Italy since April 1996 (Table A.16 of Atkinson et al. [19]). These data include two different kinds of funds (stock funds and balanced funds). From Figure 6(c), the weighted ranks contour of components 1 and 3, which explain 96.4% of the total variances, suggests there are two clusters. Moreover, the confirmatory plots provide a valid assignment of the observations, which is consistent with the two types of funds.

The third dataset, the old faithful geyser data, is taken from Azzalini and Bowman [20] and the MASS library of Venables and Ripley [21]. It includes 272 observations and two variables, the waiting time between eruptions, and the duration of the eruption in minutes for the old faithful geyser in Yellowstone National Park, Wyoming, USA. This dataset consists of two apparent clusters, the short and the long eruptions. From Figures 7(b) and 7(c), it can be seen that the weighted ranks contour of the data indicates two clusters with an unassigned observation (number 174). Using the confirmatory classifier shown in Figure 7(d), observation 174 is correctly assigned to the second cluster.

6. Comparison with Other Clustering Methods

The WSRN method determines the number of clusters in a dataset and classifies the data into each of the clusters. In this section, we compare the WSRN method with other clustering and classification methods.

The first method is the model-based clustering "mclust" [22]. This is based on a Gaussian mixture model GMM [23] where the number of clusters corresponds to the model which returns the largest Bayesian information criterion (BIC). The second method is the K-means algorithm combined with the Calinski–Harabasz (CH) index [24]. The number of clusters that returns the highest CH index is selected before applying the K-means [25] algorithm to classify the data.

The third method used as a comparator is the high-dimensional data clustering (HDDC) [26] which is again a clustering method based on the Gaussian mixture model where the BIC is used to select the number of clusters. The fourth method used is the mixture of probabilistic principal component analyses "MixtPPCA" [27] where the number of clusters corresponds to the largest BIC. The fifth method for comparison is the partitioning around medoids "PAM" clustering [28] method, where the number of clusters is estimated based on the optimum average silhouette width [29]. The sixth method used in the comparison is the density-based spatial clustering of applications with noise (DBSCAN), where the number of clusters is estimated using a density-based approach to identify regions of high density in the data and these are considered as clusters [1]. Other methods that have been used in the comparison are KMD: clustering with K-medoids [7], FCM: fuzzy C-means clustering [30], GG: Gath–Geva clustering algorithm [31], DDC: distance density clustering [8], SNN: clustering with shared nearest neighbor clustering [32], and densityClust: clustering by fast search and find of density peaks [9].

Each method was applied to the three real datasets in Section 5. As the external classes are known, the different clustering methods were compared using the purity, entropy, and the misclassification rate. Although the purity and entropy are external validation methods commonly used in classification, they measure the homogeneity of the data in clusters and do not penalize algorithms that identify the incorrect number of clusters. Indeed if each cluster is homogeneous for a particular class, both the purity and entropy will assign a perfect score (1 for purity, 0 for entropy) even if the number of clusters is incorrect. The following misclassification rate does penalize algorithms which identify the incorrect number of clusters.

Let there be n data points where there are r true classes such that $T = \{T_1, T_2, \ldots, T_r\}$ and in which the algorithm identifies k clusters such that $C = \{C_1, C_2, \ldots, C_k\}$. Let $A = \{1, 2, \ldots, k\}$ and $B = \{1, 2, \ldots, r\}$. The misclassification rate H is defined as

$$H = 1 - \left(\frac{1}{n}\right) \max\left\{ \sum_{(i,j)\in A\times B} \left|C_i \cap T_j\right| \right\}, \tag{15}$$

subject to the constraint that if two terms $|C_i \cap T_j|$ and $|C_t \cap T_u|$ appear in the sum, then $i = t$ if and only if $j = u$ [14, 33].

Thus, each row and column of the matrix $A \times B$ contribute at most one element to the sum. A consequence of this is that $|C_i \cap T_u|$ is set to zero if $|C_i \cap T_j|$ is one of the terms that maximises the sum in parentheses. Also, when there is only one cluster, the sum contains one term only.

The adjusted Rand index (ARI) is another commonly used metric for evaluating the performance of clustering algorithms [34]. While H compares clusters based on set matching, ARI assesses clusters by counting point pairs where there is agreement or disagreement. Moreover, ARI takes into account the expected value of the unadjusted Rand index, which is determined by randomly selecting entries in the contingency table with fixed column and row totals.

Other cluster validity indices can be used to evaluate the goodness of the different clustering structures such as the

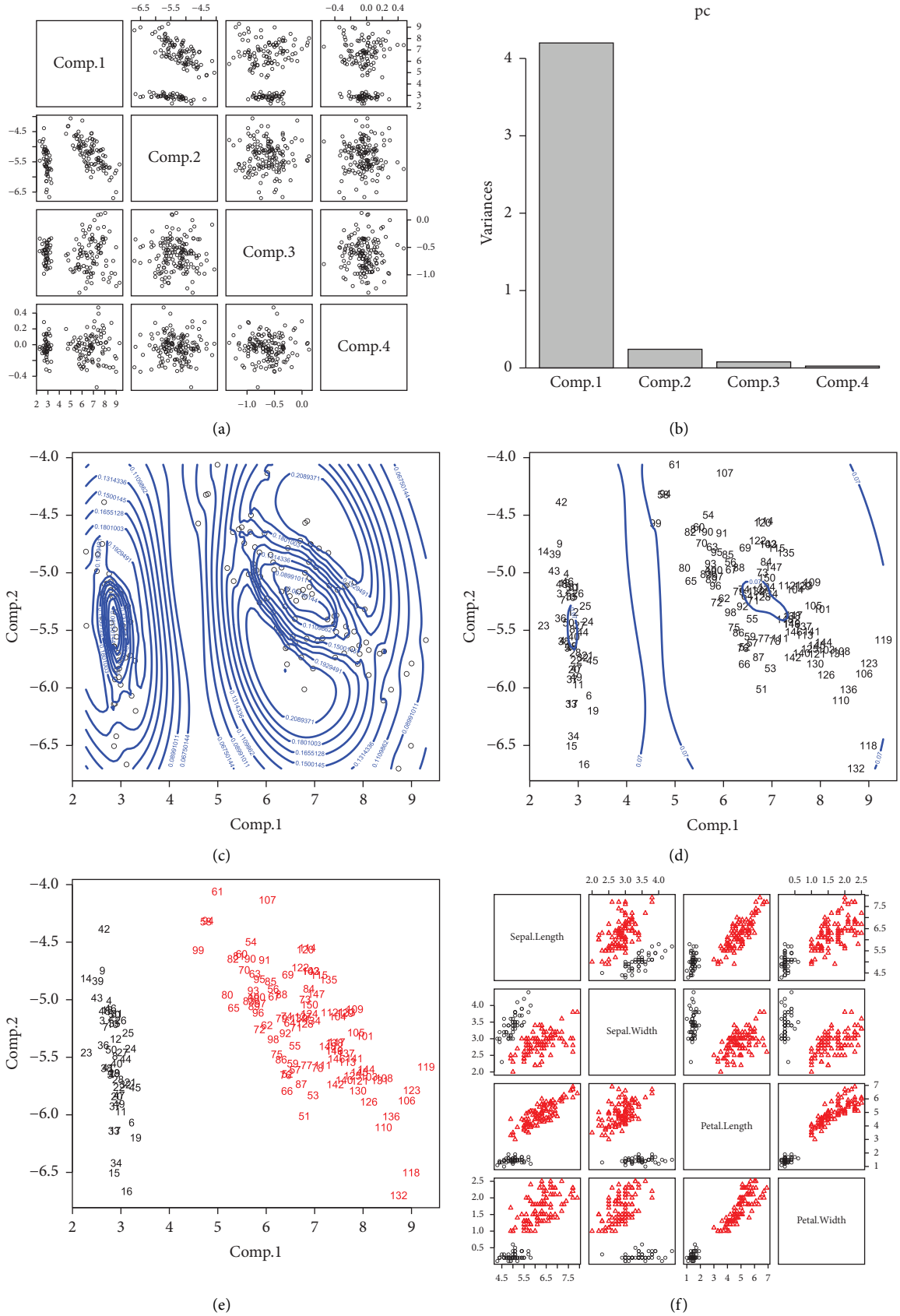


FIGURE 5: Iris data: (a) scatterplot matrix of the PCA components, (b) the total variance explained by each component, (c) the normed weighted spatial ranks contour, (d) the contour at level 0.07 and the confirmatory plots based on weighted ranks classifier for (e) the first 2 components and (f) the original data.

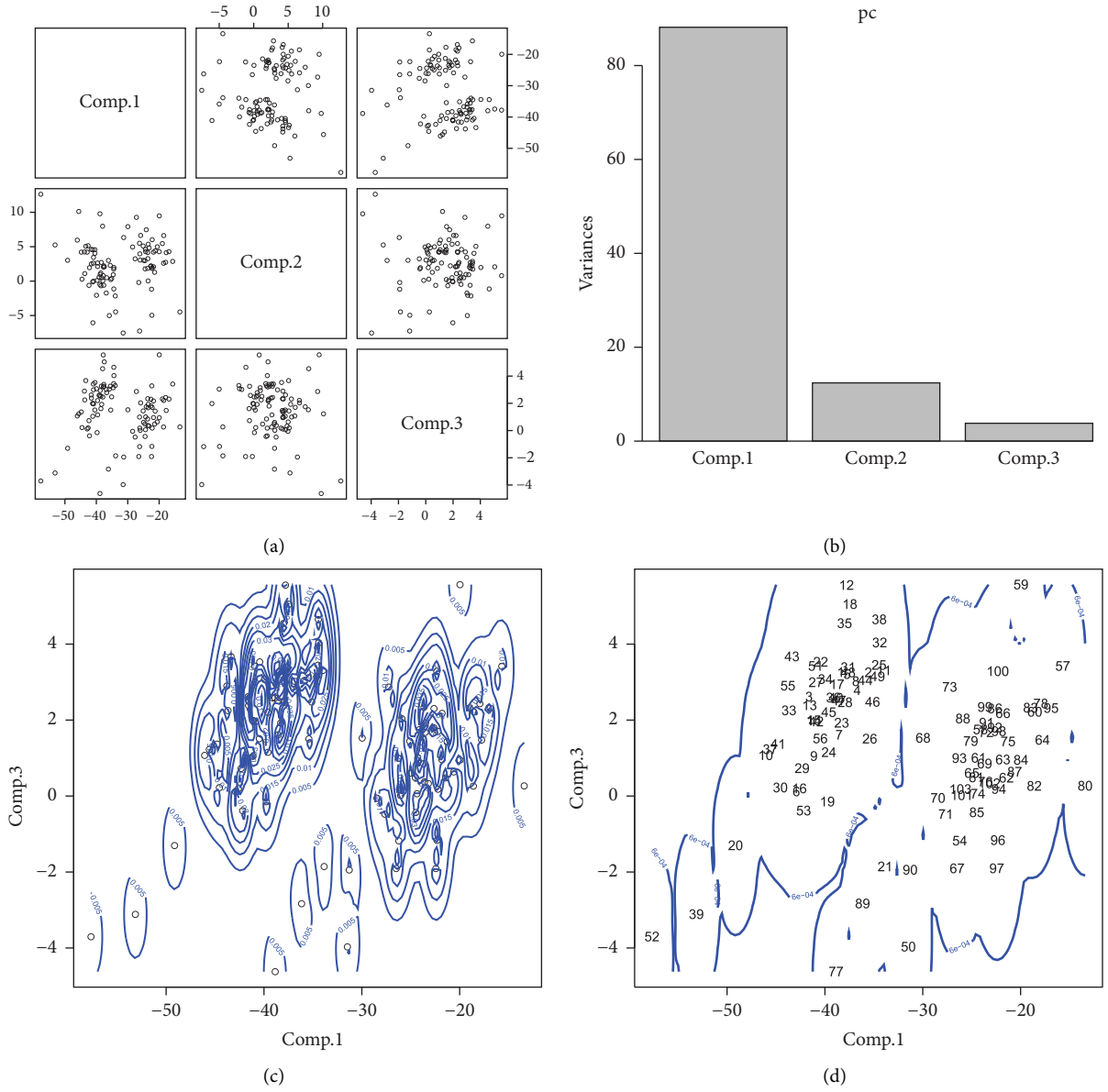
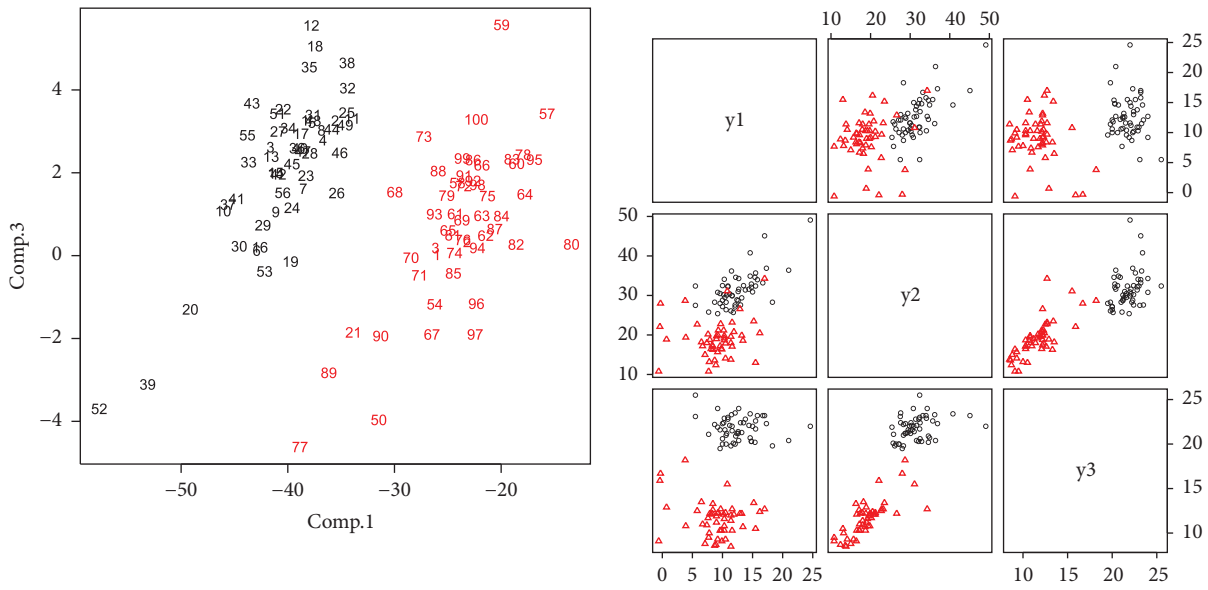
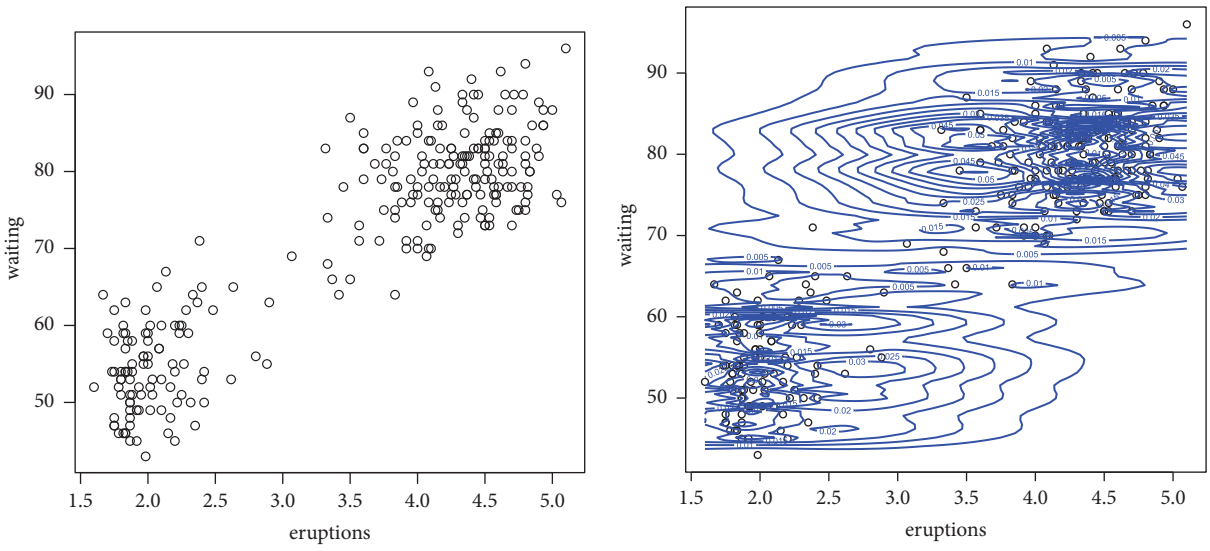


Figure 6: Continued.



(e) (f)

FIGURE 6: Financial data: (a) scatterplot matrix of the PCA components, (b) the total variance explained by each component, (c) the normed weighted spatial ranks contour, (d) the contour at level 0.0006 and the confirmatory plots based on weighted ranks classifier for (e) the first 2 components and (f) the original data.



(a) (b)

FIGURE 7: Continued.

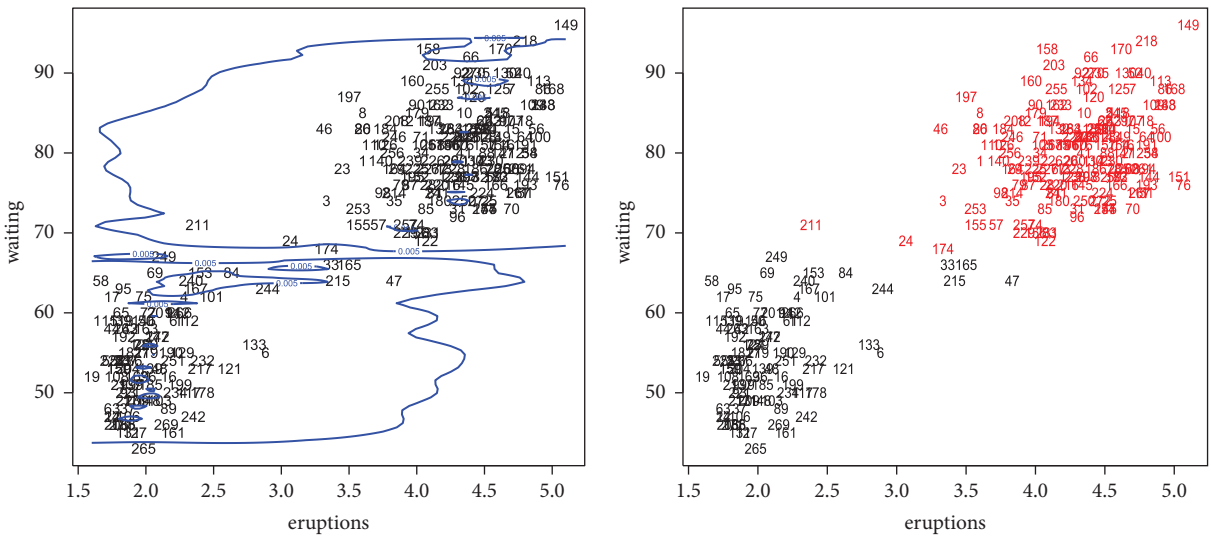


Figure 7: Old faithful data: (a) scatterplot of the faithful data, (b) the normed weighted spatial ranks contour, (c) the contour at level 0.005, and (d) the confirmatory plots based on weighted ranks classifier for original data.

Table 1: Comparison of different clustering approaches applied to the iris dataset.

Clustering method	No. of clusters	Cluster sizes	H	Purity	Entropy	ARI
WSR	2	50, 100	0	1	0	1
GMM (mclust) "BIC"	2	50, 100	0	1	0	1
K-means "CH index"*	3	38, 62, 50	0.25	1	0	0.59
HDDC "BIC"*	3	45, 55, 50	0.24	1	0	0.65
MixtPPCA "BIC"	3	50, 52, 48	0.32	1	0	0.57
PAM "silhouette width"	2	51, 99	0.01	0.99	0.05	0.97
DBSCAN	2	50, 97 (3 noise points)	0.02	1	0	0.5
KMD	3	50, 42, 58	0.28	1	0	0.58
FCM	3	60, 50, 40	0.27	1	0	0.58
GG	3	50, 65, 35	0.23	1	0	0.61
DDC	3	57, 57, 36	0.29	0.95	0.20	0.47
SNN	3	47, 94, 9	0.06	0.98	0.06	0.87
densityClust	2	50, 100	0	1	0	1

*Results are based on the mean of 1000 repetitions.

Table 2: Comparison of different clustering approaches applied to the financial dataset.

Clustering method	No. of clusters	Cluster sizes	H	Purity	Entropy	ARI
WSR	2	53, 50	0.03	0.97	0.16	0.89
GMM (mclust) "BIC"	3	50, 15, 38	0.15	0.94	0.14	0.73
K-means "CH index"*	2	46, 57	0.03	0.97	0.19	0.89
HDDC "BIC"*	3	8, 45, 50	0.09	0.98	0.09	0.82
MixtPPCA "BIC"	2	53, 50	0.03	0.97	0.19	0.89
PAM "silhouette width"	2	57, 46	0.03	0.97	0.19	0.89
DBSCAN	1	102 (1 noise point)	0.03	0.54	0.99	−0.003
KMD	2	46, 57	0.03	0.97	0.19	0.89
FCM	2	46, 57	0.03	0.97	0.19	0.89
GG	2	53, 50	0.03	0.97	0.16	0.89
DDC	3	11, 46, 46	0.12	0.97	0.14	0.76
SNN	3	47, 40, 16	0.16	0.93	0.15	0.71
densityClust	1	103	0	0.54	0.99	0

*Results are based on the mean of 1000 repetitions.

Table 3: Comparison of different clustering approaches applied to the old faithful dataset.

Clustering method	No. of clusters	Cluster sizes	H	Purity	Entropy	ARI
WSR	2	172, 100	0.018	0.98	0.12	0.93
GMM (mclust) "BIC"	3	40, 97, 135	0.15	1	0	0.71
K-means "CH index"*	10	46, 42, 19, 23, 48, 21, 23, 13, 20, 17	0.71	0.98	0.07	0.18
HDDC "BIC"*	2	175, 97	0	1	0	1
MixtPPCA "BIC"	2	174, 98	0.004	0.996	0.03	0.98
PAM "silhouette width"	2	172, 100	0.018	0.98	0.12	0.93
DBSCAN	3	168, 82, 17 (5 noise points)	0.08	0.98	0.09	0.85
KMD	3	97, 76, 99	0.28	1	0	0.60
FCM	2	172, 100	0.018	0.98	0.12	0.93
GG	2	97, 175	0	1	0	1
DDC	2	172, 100	0.018	0.98	0.12	0.93
SNN	2	270, 2	0.01	0.64	0.94	−0.01
densityClust	1	272	0	0.64	0.94	0

*Results are based on the mean of 1000 repetitions.

connectivity index [35], CS index [36], and Sym index [37] (for more extensive details, see [38]).

Table 1 shows the results of the different algorithms applied to the iris dataset, where nine of the twelve methods recorded perfect scores for the purity and entropy, despite only 4 algorithms identifying the correct number of clusters. For these data, the WSR, mclust, and densityClust have the best misclassification rate, and H and both have perfect entropy, purity, and ARI scores.

For the financial dataset, as shown in Table 2, the WSR with seven other algorithms have the joint lowest misclassification rates. The HDDC algorithm records the best scores for the purity and entropy but this identifies an incorrect number of clusters. Based on the ARI, WSR, K-means, MixtPPCA, PAM, KMD, FCM, and GG record the best scores. Finally, for the old faithful dataset, Table 3 shows that the WSR algorithm has the joint third lowest H, but the HDDC and GG algorithms are the best across all the four metrics for this dataset.

7. Concluding Remarks

In this paper, we have introduced a new clustering method based on weighted spatial ranks. The WSRN algorithm is completely data-driven and it both determines the number of clusters and classifies the data. As a nonparametric method, it does not require any assumptions to be made on the underlying distribution(s) of the data. The synthesis of weighted rank contours, based on principal components analysis when the data have more than two dimensions, allows the intuitive visualization of the cluster structure in relation to the distribution of the data points.

We considered nonparametric kernel weights and we introduced WSRN functions based on Gaussian kernel weights. Compared to other standard approaches, the WSRN function based on the Gaussian kernel weights provided the best results in terms of cluster detection and visualization. The weighted rank contours based on Gaussian weights were more accurate and provided the best fit to the shape of the clusters' structure. They captured each observation carefully and assigned it to the proper group with a minimal probability of misclassification. It also performed competitively with other methods when clustering and classifying the data from three real datasets.

Although the WSRN method is invariant under orthogonal transformations, it is not an affine invariant. Using affine invariant ranks has the potential to improve the results if the scales of different clusters are not similar [39]. A further possible extension to the method would be to consider generalizations of the Euclidean norm for estimating the WSRN. Thus, different Lp norms for $p > 2$ could be investigated to establish the optimal value for p.

Data Availability

The data used to support the findings of the study are available in the public domain and are appropriately referenced in this article.

Disclosure

This paper is a modified version of part of the published work in MB's Thesis (Baragilly [40]).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank Biman Chakraborty for his helpful discussions and suggestions relating to this work, and BHW and MB were supported by a Clinician Scientist award with the Medical Research Council, UK (MR/N007999/1).

References

[1] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, Portland Oregon, August 1996.

[2] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, Inc, Hoboken, NY, USA, 1988.

[3] J. B. Macqueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, 1967.

[4] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," *Advances in Neural Information Processing Systems*, pp. 849–856, 2002.

[5] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[6] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, NY, USA, 1981.

[7] H. S. Park and C. H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3336–3341, 2009.

[8] R. Ma and R. Angryk, "Distance and density clustering for time series data," in *Proceedings of the 2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, IEEE, New Orleans, LA, USA, November 2017.

[9] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[10] M. Suo, B. Zhu, D. Zhou, R. An, and S. Li, "Neighborhood grid clustering and its application in fault diagnosis of satellite power system," *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, vol. 233, no. 4, pp. 1270–1283, 2019.

[11] M. Baragilly and B. Chakraborty, "Determining the number of clusters using multivariate ranks," in *Recent Advances in Robust Statistics: Theory and Applications*, C. Agostinelli, A. Basu, P. Filzmoser, and D. Mukherjee, Eds., pp. 19-36, Springer, India, 2016.

[12] J. Möttönen and H. Oja, "Multivariate spatial sign and rank methods," *Journal of Nonparametric Statistics*, vol. 5, no. 2, pp. 201–213, 1995.

[13] S. Sirkiä, S. Taskinen, H. Oja, and D. E. Tyler, "Tests and estimates of shape based on spatial signs and ranks," *Journal of Nonparametric Statistics*, vol. 21, no. 2, pp. 155–176, 2009.

[14] M. Baragilly, H. Gabr, and B. H. Willis, "Clustering functional data using forward search based on functional spatial ranks with medical applications," *Statistical Methods in Medical Research*, vol. 31, no. 1, pp. 47–61, 2021.

[15] C. R. Souza, "Kernel functions for machine learning applications," 2010, http://crsouza.blogspot.com/2010/03/kernel-functions-for-machine-learning.html.

[16] I. T. Jolliffe, *Principal Component Analysis*, Springer, Berlin, Germany, 2002.

[17] W. J. Kzanowski, *Principles of Multivariate Analysis*, OUP, Oxford, UK, 1988.

[18] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[19] A. C. Atkinson, M. Riani, and A. Cerioli, *Exploring Multivariate Data with the Forward Search*, Springer, NewYork, NY, USA, 2004.

[20] A. Azzalini and A. Bowman, "A look at some data on the old faithful geyser," *Applied Statistics*, vol. 39, no. 3, pp. 357–365, 1990.

[21] W. Venables and B. Ripley, *Modern Applied Statistics with S*, Springer, New York, NY, USA, 2002.

[22] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611–631, 2002.

[23] J. D. Banfield and A. E. Raftery, "Model-based Gaussian and non-Gaussian clustering," *Biometrics. JSTOR*, vol. 21, p. 803, 1993.

[24] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-Theory and Methods*, vol. 3, pp. 1–27, 1974.

[25] S. P. Lloyd, "Least squares quantization in PCM. Technical note, bell laboratories," *IEEE Transactions on Information Theory*, vol. 28, pp. 128–137, 1957.

[26] C. Bouveyron, S. Girard, and C. Schmid, "High dimensional data clustering," *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 502–519, 2007.

[27] M. E. Tipping and C. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[28] A. Reynolds, G. Richards, B. de la Iglesia, and V. Rayward-Smith, "Clustering rules: a comparison of partitioning and hierarchical clustering algorithms," *Journal of Mathematical Modelling and Algorithms*, vol. 5, no. 4, pp. 475–504, 2006.

[29] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data. An Introduction to Cluster Analysis*, John Wiley and Sons Inc, New York, NY, USA, 1990.

[30] A. J. Torabi, M. J. Er, X. Li, B. S. Lim, and G. O. Peen, "Application of clustering methods for online tool condition monitoring and fault diagnosis in high-speed milling processes," *IEEE Systems Journal*, vol. 10, no. 2, pp. 721–732, 2016.

[31] K. Yu, T. R. Lin, and J. W. Tan, "A bearing fault diagnosis technique based on singular values of EEMD spatial condition matrix and Gath-Geva clustering," *Applied Acoustics*, vol. 121, pp. 33–45, 2017.

[32] L. Ertoz, M. Steinbach, and V. Kumar, "Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data," in *Proceedings of the SIAM International Conference on Data Mining*, pp. 47–59, Minnesota, MI, USA, April 2003.

[33] M. Meila, "Comparing clusterings—an information based distance," *Journal of Multivariate Analysis*, vol. 98, no. 5, pp. 873–895, 2007.

[34] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[35] J. Handl and J. Knowles, "An evolutionary approach to multiobjective clustering," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 1, pp. 56–76, 2007.

[36] C. H. Chou, M. C. Su, and E. Lai, "A new cluster validity measure and its application to image compression," *Pattern Analysis & Applications*, vol. 7, no. 2, pp. 205–220, 2004.

[37] S. Bandyopadhyay and S. Saha, "A point symmetry-based clustering technique for automatic evolution of clusters,"

IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 11, pp. 1441–1457, 2008.

[38] A. José-García and W. Gómez-Flores, "Automatic clustering using nature-inspired metaheuristics: a survey," *Applied Soft Computing*, vol. 41, pp. 192–213, 2016.

[39] B. Chakraborty, "On affine equivariant multivariate quantiles," *Annals of the Institute of Statistical Mathematics*, vol. 53, no. 2, pp. 380–403, 2001.

[40] M. H. H. Baragilly, *Clustering Multivariate and Functional Data Using Spatial Rank Functions*, University of Birmingham, Birmingham, UK, 2016.