Circular Clustering with Polar Coordinate Reconstruction

Xiaoxiao Sun, *Member, IEEE,* and Paul Sajda, *Fellow, IEEE*

Abstract—There is a growing interest in characterizing circular data found in biological systems. Such data are wide-ranging and varied, from the signal phase in neural recordings to nucleotide sequences in round genomes. Traditional clustering algorithms are often inadequate due to their limited ability to distinguish differences in the periodic component θ. Current clustering schemes for polar coordinate systems have limitations, such as being only angle-focused or lacking generality. To overcome these limitations, we propose a new analysis framework that utilizes projections onto a cylindrical coordinate system to represent objects in a polar coordinate system optimally. Using the mathematical properties of circular data, we show that our approach always finds the correct clustering result within the reconstructed dataset, given sufficient periodic repetitions of the data. This framework is generally applicable and adaptable to most state-of-the-art clustering algorithms. We demonstrate on synthetic and real data that our method generates more appropriate and consistent clustering results than standard methods. In summary, our proposed analysis framework overcomes the limitations of existing polar coordinate-based clustering methods and provides an accurate and efficient way to cluster circular data. We provide the code as open-source on GitHub.

Index Terms—Circular clustering, Polar Coordinate, Coordinate Reconstruction, Phase Synchronization, Circular DNA molecules

1 Introduction

Clustering is a common unsupervised learning technique for statistical analysis used to explore interesting patterns inside large datasets [1]. It has been generally applied in many fields, including machine learning, data mining, pattern recognition, image analysis, and bioinformatics [2]. During the past decade, a number of clustering algorithms have been proposed and categorized based on their different characteristics (e.g., partitioning, hierarchical, density-based, etc.) [2, 3]. Corresponding distance measures are used to determine the similarity/dissimilarity between objects, with differences in these measures potentially resulting in significantly different and inconsistent clustering outcomes [4].

In terms of application, there is a growing interest in analyzing circular data from different types of biological systems. For instance, the rhythmically synchronized activity in the brain is widely believed to play a fundamental role in coordinating information transfer during goal-directed behavior [5, 6]. Recent studies suggest that the phase of ongoing oscillatory activity relative to endogenous or exogenous cues will facilitate coordinated information transfer within circuits and between distributed brain areas and provide windows for optimal communication between two or more brain areas [7, 8, 9]. Investigating phase synchronization, which is represented with polar coordinates, requires appropriate analysis tools to characterize aspects of the oscillatory activity, and one approach to do this is via unsupervised clustering. Another example is the analysis of round genomes that widely exist in living systems, including circular RNA and extrachromosomal circular DNA molecules. These have been linked to multiple diseases including cancer [10, 11], and thus the clustering of CpG islands, gene locations and origin of replication can help discover active regions or hot spots [12, 13, 14]. Additional applications of circular clustering can also be found in iridology for disease diagnosis [15, 16]. Lately, the investigation of cluster analysis methods applied to protein sequences employing polar coordinate representations has also garnered increasing attention and exploration [17, 18, 19].

In considering how clustering methods can be applied to oscillatory/circular data, it is important to note that most state-of-the-art clustering algorithms are not constructed in a way that is based on a polar coordinate representation. Although there are some circular clustering tools, they either only emphasize a polar angle difference (ignore the amplitude) or are designed with loss of generality – i.e., they are designed to work on datasets that obey specific assumptions [2, 3, 14]. Due to the absence of suitable tools, certain studies may resort to sub-optimal approaches [20] or concentrate exclusively on the angle difference, disregarding the radius aspect [21, 22].

In this paper, we propose a novel method that results in a polar coordinate representation that is directly applicable to most cluster methods that are not specifically designed to analyze circular data. Specifically, our approach maps polar coordinates to the lateral surface of cylindrical coordinates, creating a 3-dimensional representation that records differences in both distance r from the origin and the angle θ. This representation is then unraveled and flattened into a rectangular plane which simplifies and generalizes the geodesic, in 3-dimensional space, to a Euclidean distance between two points in a 2-dimensional space, thereby avoiding the need

- *P. Sajda is with Department of Biomedical Engineering, Electrical Engineering, Radiology, and Data Science Institute, Columbia University, New York, NY, 10027. E-mail: psajda@columbia.edu*
- *X. Sun is with Department of Biomedical Engineering, Columbia University, New York, NY, 10027. E-mail: xiaoxiao.sun@columbia.edu*

for complex algorithm design and computation to obtain the 3-d distance metric. This reconstructed coordinate can then be directly used as an input to most clustering algorithms. We address the issue of circularity/periodicity in data by searching for clusters across multiple identical periods with repetitions of the original period. We present, as examples, methods that apply to partitioning (e.g., K-means), density-based (e.g., Density-Based Spatial Clustering of Applications with Noise (DBSCAN)), and hierarchical clustering (e.g., dendrogram) methods. Lastly, we demonstrate the utility of using this clustering approach for synthetic data and real biological data (e.g., phase synchronization of neural data and circular DNA molecules).

2 Related Work

The main objective of the circular clustering algorithm is to divide data points into groups or clusters arranged in a circular or ring-shaped pattern [2, 3]. The circular pattern can be particularly useful when the data is arranged circularly or periodically, such as in the analysis of time series data, biological data, or geographic data [4, 23].

One of the earliest and most well-known circular clustering algorithms is the K-means algorithm. This algorithm partitions the data into a pre-specified number of clusters, representing each cluster by its centroid [13]. While K-means is a powerful clustering algorithm, it has some limitations when it comes to circular clustering. For example, K-means usually uses Cartesian distance as its distance metric, which may not be appropriate for circular data [23, 24].

To overcome these limitations, several variations of the K-means algorithm have been proposed [14, 23, 24]. For instance, the K-means clustering algorithm for circular data (KCC) uses a polar coordinate system to represent the data, where the distance between two data points is measured along the circumference of a circle [23]. The algorithm then adjusts the centroids based on the average angle and radius of the data points in each cluster. Circular K-means (CK-means) is another K-means based algorithm where it creates cluster vectors to contain directional information in a circular-shift invariant manner [24]. Furthermore, traditional single-objective clustering algorithms relying on intracluster distance (ICD) functions prove to be unsuitable for data that is not well-separated [25]. To address this, an accelerated two-stage particle swarm optimization (ATPSO) method is introduced, which incorporates K-means clustering to expedite particle convergence during population initialization, leading to more precise and rapid cluster detection [26].

Despite their usefulness, K-means based algorithms have certain limitations, such as their sensitivity to initialization, the need for prior knowledge of the number of clusters, and difficulties with unequal cluster sizes. As a result, alternative circular clustering algorithms that do not rely on K-means have been developed, including mean shift-based [27] and learning-based clustering methods [28, 29, 30]. However, it is important to note that the performance of mean shift-based methods can be influenced by the bandwidth parameter of the circular kernel function, which controls the smoothing of the density function. Moreover, learning-based clustering methods require labeled data for training, which can be expensive and time-consuming to obtain [31]. In addition, these methods perform sub-optimally when confronted with complex data distributions that are not consistent with the model assumptions. Lastly, as the models employed can be quite complex with numerous parameters, interpretation of the results can be difficult.

Alternative approaches for handling nonlinear data include spectral clustering, which utilizes the spectral properties of the data's affinity matrix. Multiple kernel clustering (MKC) methods, in particular, widely adopt this approach [32, 33]. These methods involve learning an affinity graph and subsequently applying spectral clustering to yield clustering results. However, like many other methods, the majority of algorithms are constructed around a Cartesian system for distance metric computation, neglecting the potential presence of circular or periodic patterns.

Overall, circular clustering is a valuable approach for finding patterns in circular data, though the existing methods have significant limitations that make them unsuitable for many applications. Therefore, there is a need for novel and versatile algorithms that can model circular patterns explicitly across a broad range of applications. Such algorithms would help to overcome the limitations of current circular clustering methods and enable more accurate and efficient analysis of circular data.

3 Methods

3.1 Distance Metrics on Cartesian Coordinates

Similarity or dissimilarity measurements in clustering algorithms play an essential role in grouping or separating data [4]. A key step is to select suitable distance metrics. The simplest measure is the Manhattan distance (i.e., L^1 norm) which is equal to the sum of absolute distances for each variable. The formula for this grid-like path distance D between $X = (x_1, x_2 \cdots, x_n)$ and $Y = (y_1, y_2 \cdots, y_n) \in \mathbb{R}^n$ is:

$$D(X, Y) = \sum_{i=1}^{n} |x_i - y_i| \tag{1}$$

A more commonly used metric in many clustering algorithms is the Euclidean distance (i.e., L^2 norm or Euclidean norm) [2]. It is computed by taking the square root of the sum of the squares of the differences between each variable. The formula for this, as-the-crow-flies, distance D between X and Y is:

$$D(X, Y) = \Big(\sum_{i=1}^{n} (|x_i - y_i|)^2 \Big)^{\frac{1}{2}} \tag{2}$$

3.2 Clustering on Polar Coordinates

A general approach for solving clustering problems in polar coordinates is to convert the representation of the variables from polar coordinates (r, θ) to Cartesian coordinates (x, y) (see (3) and Fig. 1**A**) and calculate the Euclidean distance with Cartesian coordinates.

$$(x, y) = (r\cos(\theta), r\sin(\theta)). \tag{3}$$

However, since polar coordinates have a circular element θ, clustering methods using the Euclidean distance can lead to

This article has been accepted for publication in IEEE/ACM Transactions on Computational Biology and Bioinformatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2024.3406341

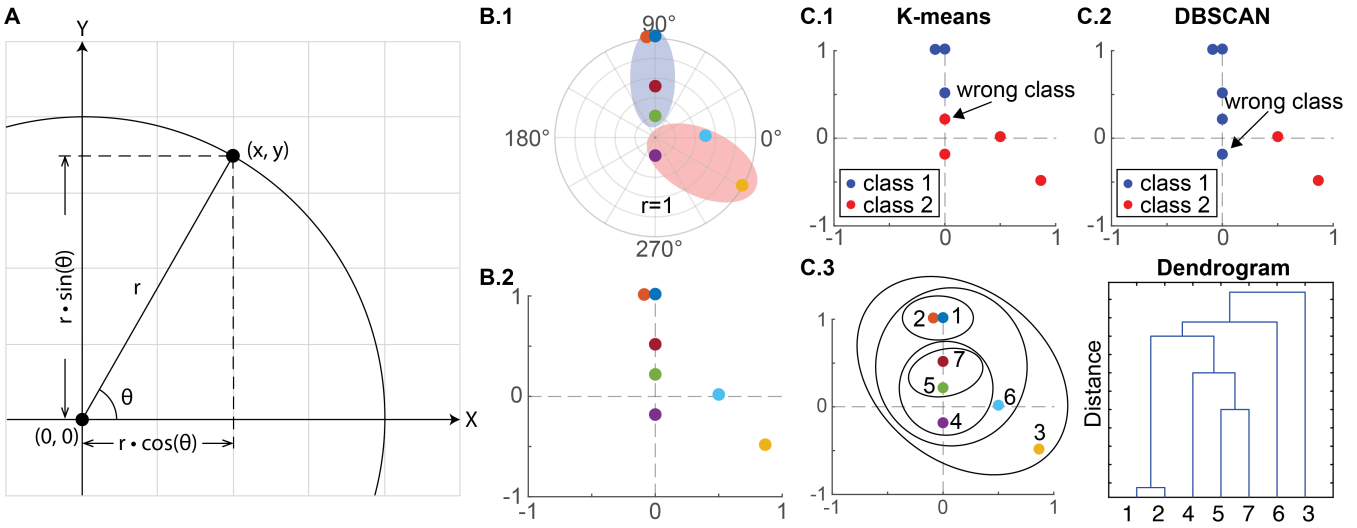

Fig. 1. Common approach to perform clustering analysis for data specified in polar coordinates. **A.** Illustration converting from polar coordinates (r, θ) to Cartesian coordinates (x,y) as (3). **B.1.** Seven example sample points in polar coordinates are shown for illustration purposes (points in the same shaded area belong to the same class). **B.2.** Transforming points in **B.1** to Cartesian coordinates (points correspondences indicated by color). **C.1.** Clustering result generated by K-means method ($K = 2$) where the green point $(r = 0.2, \theta = \pi/2)$ in **B.1** is incorrectly classified. **C.2.** Clustering result generated by DBSCAN method (neighborhood search radius $\epsilon = 0.5$ and minimum number of neighbors $n_{\min} = 2$) where the purple point $(r = 0.2, \theta = 3\pi/2)$ in **B.1** is incorrectly classified. **C.3.** Clustering result generated by the hierarchical clustering method. The figure on the left graphically illustrates the way the linkages group the objects into a hierarchy of clusters, and the figure on the right is the corresponding dendrogram of clusters where the x-axis corresponds to the leaf nodes of the tree, and the y-axis corresponds to the linkage distances between clusters. Similar to the K-means and DBSCAN methods, incorrect grouping arises for the sample points that are labeled as 4, 5, and 7.

an incorrect result. For instance, if two points have a $180°$ phase difference but are close to the origin, any L^2-based clustering method would give a wrong conclusion because of its limited ability to distinguish differences in θ (i.e., these two misclassified points with the same radius but $180°$ out-of-phase in Fig. 1**B.1** can also be interpreted as in-phase but having opposite amplitudes in Fig. 1**B.2**). Furthermore, a clustering method with Cartesian coordinates may not be able to extract corresponding meaningful information in polar coordinates. While several circular clustering algorithms have previously been proposed to address these issues, most of them only differentiate the θ component [2, 14] or are extremely complex [3, 29], leaving researchers with no direct or generally applicable clustering tools optimized for data in polar coordinates.

3.3 Polar Coordinate Reconstruction

Since Euclidean distance is not a suitable metric for data in polar coordinates, one straightforward solution is to use other distance metrics (e.g., ArcCosineDistance, Canberra Distance, Lorentzian Distance, etc.) [2, 4]. However, this requires researchers to recreate different clustering algorithms specialized for their data, though this results in a loss of generality. Moreover, such methods may still fail to capture the differences in both r and θ (e.g., most circular clustering algorithms emphasize θ [14]). Alternatively, one can convert a polar representation of points to a non-cartesian representation, though it is challenging to deal with the periodic element ($\theta = \theta + 2k\pi, k \in \mathbb{N}$) and its circular characteristic ($\Delta\theta = \theta_1 - \theta_2$ or $\Delta\theta = \theta_2 - \theta_1$, i.e., the angle difference between two points can be approached either clockwise or counterclockwise). Additionally, these converted coordinates may not be well-suited for state-of-the-art clustering methods. Thus, there is a need for a circular clustering framework that is generalizable across clustering approaches and enables one to consider both magnitude and phase angle in the distance metric.

3.3.1 *Proposed Coordinate System*

Since cylindrical coordinates are an extension of 2-dimensional polar coordinates to 3 dimensions, we propose to use such a coordinate system to capture dissimilarities in both r and θ. After mapping points in polar coordinates to the lateral surface of the cylinder with (4), the distance between two points on the lateral surface (i.e., geodesic) can be estimated as the Euclidean distance on the rectangular space (X', Y'). This rectangular space is obtained by unraveling the cylinder and flattening it into a plane (see Fig. 2, where the lateral surface area of a cylinder is the area of a rectangle). Any point (x', y') on (X', Y') can be written as (5),

$$(x, y, z) = (R\cos(\theta), R\sin(\theta), r) \tag{4}$$

$$(x', y') = (f(\theta), r) = (R\theta, r) \tag{5}$$

where R is the radius of the base circle of the cylinder. R is a parameter that can be adjusted based on the weights (importance) of difference in $r \in [0, \infty)$ and $\theta \in [0, 2\pi)$. Greater R indicates the difference observed in θ is more important than r.

3.3.2 *Solving for Circularity*

As shown in the rectangular space of Fig. 2, because point B on the cylindrical coordinates can be expressed as either $B1 = (0, r)$ or $B2 = (2\pi, r)$, then the distance between example point A and B is given by either $D1(A, B1)$ or

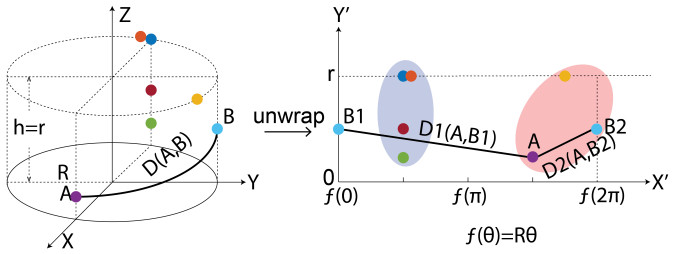

Fig. 2. Reconstruction of polar coordinates on the rectangular plane (X', Y') using cylindrical coordinates. Mapping the points in polar coordinates to the lateral surface of the cylinder where the height is equal to the radius of the polar coordinates system ($h = r$) and the base circle has radius R. Values of R control the weights of θ as described in (5). By flattening the lateral surface, we obtain a 2-dimensional rectangular space where the X'-axis represents θ and Y'-axis represents r. The length of the X'-axis equals the circumference of the base circle (i.e., $C = 2\pi R$).

$D2(A, B2)$. In this example, the appropriate distance to generate the correct clustering result is $D2(A, B2)$, so it is essential to locate the accurate distance that moves clockwise or counterclockwise in a circle. Inspired by the coordinate optimization approach in phase-controlled robotics [34], we can take advantage of the periodic nature of the polar coordinate system. Specifically, we address the circularity problem by leveraging its periodic nature through period repetition and explore three different clustering methods to demonstrate how this technique can be used.

Period Repetition: As mentioned previously, obtaining the correct minimal distance is key to identifying the accurate clustering outcomes, so here we will discuss how this distance can be revealed by period repetition. We will demonstrate its validity through a simplified example, and more complex scenarios can be proven using mathematical induction.

1) Within-cluster variance is unchanged after period repetition (see Fig. 3**A**): In the original space (period = 1), given a dataset with three sample points $\{X_{1,1}, X_{2,1}, X_{3,1}\}$ (where $X_{i,n}$ indicates the i-th data point in period n), the goal of classifying the data into 2 clusters with a minimal sum of distances within each cluster is equivalent of finding the two data points with the smallest distance to form one class and the point left would be the other class (by Triangle Inequality Theorem which states that in a Euclidean space, the shortest distance between two points is a straight line). Then two cluster centroids would be μ_1 and μ_2 where μ_1 is the middle point between the smallest distance and μ_2 is the left point itself. By repeating two periods ($n = 2$), the problem becomes finding 4 (i.e., $2 \times n$) classes from the given dataset $\{X_{1,1}, X_{2,1}, X_{3,1}, X_{1,2}, X_{2,2}, X_{3,2}\}$. Because the between period point distance is much greater than any within period point distance (e.g., $d(X_{3,1}, X_{2,2}) > d(X_{1,1}, X_{2,2}) \gg d(X_{i,n}, X_{j,n}), n \in \{1, 2\}$), the new formed cluster centroids $\{\mu_1, \mu_2\}$ and $\{\mu_3, \mu_4\}$ would follow similar pattern as $\{\mu_1, \mu_2\}$ in original space – within cluster variance is not unchanged after the repetition. Consequentially, the clustering result is identical as the original space, which means class $(c1, c2)$ formed in original space is the same as $(c1, c2)$ and $(c3, c4)$ formed in the repeated space. This conclusion can be generalized to the scenario where $n = k$ and $i = k$.

2) Within-cluster variance is changed after period repetition (see Fig. 3**B**): similar as the aforementioned example, the goal is to classify the given dataset with three sample points $\{X_{1,1}, X_{2,1}, X_{3,1}\}$ into 2 clusters. In the original space, because of the misrepresentation of the distance between X_1 and X_2, it generates the wrong conclusion compared to the ground truth ($d(X_{1,1}, X_{2,1}) > d(X_{1,1}, X_{3,1})$). By repeating three periods ($n = 3$, repeating an odd number of periods makes the left and right of the middle period balanced to simplify the searching process within the middle period), the correct minimal distance measured between X_1 and X_2 is revealed ($d(X_{1,1}, X_{2,2}) < d(X_{1,1}, X_{3,1}) < d(X_{1,1}, X_{2,1})$). Additionally, since the within-cluster variance is changed, this appearance of $d(X_{1,1}, X_{2,2})$ forms a new type of centroids (μ_3 and μ_5), which leads to the correct clustering results as ground truth – i.e., $c3$ and $c4$ is the accurate clustering outcomes given dataset $\{X_1, X_2, X_3\}$. With greater number of period repetitions ($n = 5$), this correct pattern would occur periodically and be identified repetitively (highlighted with the dashed red rectangular box in Fig. 3**B**).

In conclusion, period repetition helps the clustering algorithms to use the accurate distance between given sample points to generate the correct centroids. A Greater number of repetitions makes the correct pattern occur robustly (i.e., periodically) and accounts for the potential randomness in initialization. To get the correct clustering results, we propose different search algorithms to identify them for different clustering methods (see the complexity analysis of search algorithms in Appendix C).

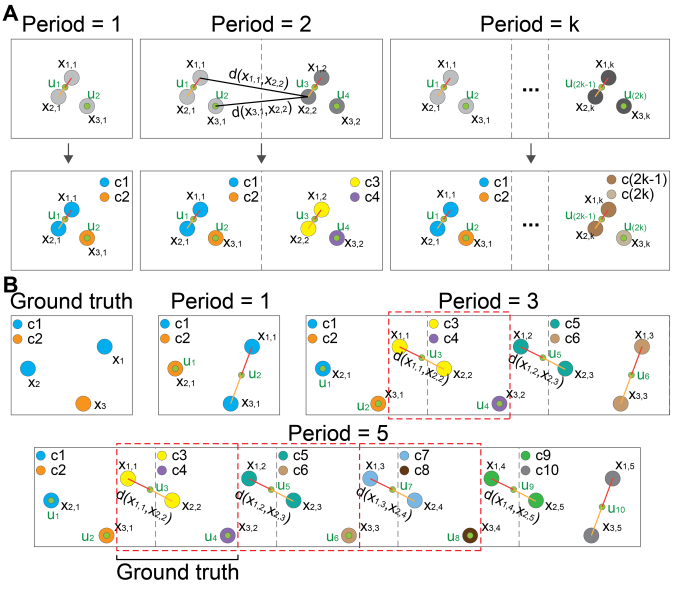

Fig. 3. Period repetition for revealing the correct distance between sample points. **A.** Within-clustering variance is unchanged after period repetition, so the clustering result within each repeated period is identical to the original space. No changes will be introduced by period repetition. Clustering results always match with the ground truth. **B.** Within-clustering variance is changed after period repetition, leading to the emergence of novel centroid types. This iterative process helps the clustering algorithm to capture the accurate distance between individual samples. As the dataset size or clusters increase, more repetitions are necessary to ensure the correct patterns occur robustly.

K-means: Because of the periodicity of points in polar coordinates, the layout of points within each extended period (from 0 to 2π) in the rectangular plane would be identical. Suppose the number of classes is predetermined as K. In that case, the question of finding K classes within one period is equivalent to the question to find $(c+1) \times K$ classes in the total extended rectangular plane when we repeat the period an additional c times (i.e., $c+1$ periods in total). Based on the similarity of the extended periods, one can always find the correct clustering results to the problem in the middle period if we repeat the periods enough times (see Fig. 4**A.1** and Algorithm 1. Note: one repeats an even number of periods to keep the data distribution around the middle periods nearly identical, e.g., $c = 2n, n = 1, 2, \cdots$).

Algorithm 1 Search algorithm for K-means

Require: $c = 2n, n \in \mathbb{Z}^+$, clustering class K, clustering result of $2n+1$ periods K_t
Ensure: Final clustering result identified K_c
locate middle period: $N \leftarrow n+1$
sample points within N: $P_{n+1} \in 360° \times [n, n+1]$
corresponding clustering class for P_{n+1}: $K_{n+1} \subset K_t$
combination of K_{n+1} based on number K: C_K
for $x = 1$ to size(C_K) **do**
    if combination C_x includes all the original sample points P_1 without repetition **then**
        $K_c \leftarrow C_x$
    end if
end for
if No such C_x exists **then**
    increase the repetition number c
end if

DBSCAN: Repeating periods also work when the number of classes is unknown (e.g., clustering with density-based methods). Adapting the same parameters (ϵ and $n_{\min}$) used for the original period, we repetitively classify points belonging to the same group into one group (see Fig. 4**A.2**). Those points within those identical classes must be classified into one group (see Algorithm 2). For example, in Fig. 4**A.2**, one type of class has been repetitively found. Another instance with different parameters is shown in Fig. 4**A.3** and Fig. 4**B.3**, where two classes with no outliers are identified. Fig. 4**B.3** provides the same conclusion as K-means method when $K = 2$ in Fig. 4**B.1**. Period repetition does not change the definition of outliers in the DBSCAN algorithm, instead, they are completely determined by the choice of ϵ and $n_{\min}$ (i.e., the default parameters). More repetitions will provide higher confidence in the clustering result since the 'correct' pattern would always be detected after a certain number of repetitions.

Hierarchical Clustering (dendrogram): Period repetition can also be used in the hierarchical clustering method to find the correct minimal pairwise distance. For example, in the given samples in Fig. 2, the distance between A and B would be D_2 which will be calculated when we repeat B_1 as B_2 in the second period, and the final hierarchical clustering result with reconstructed coordinates is given in Fig. 4**B.4**. Similarly, repetition is needed to reveal all the possible pairwise distances. The details of this algorithm can be found in Algorithm 3.

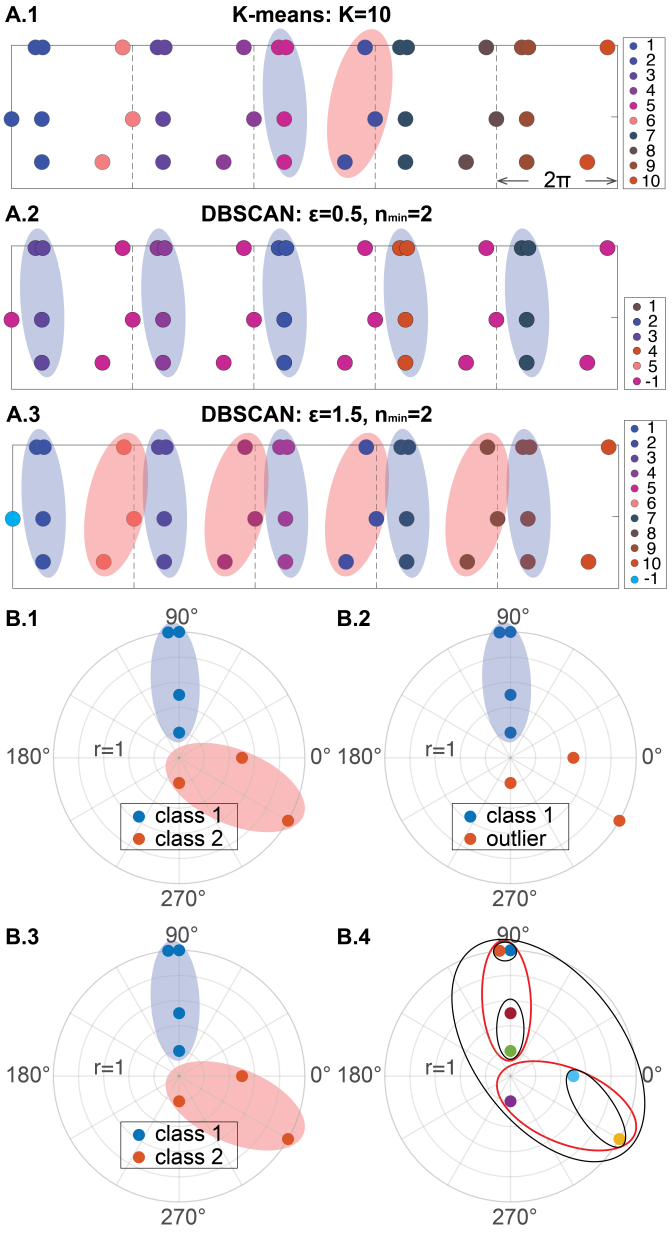

Fig. 4. Solving circularity with period repetition for K-means, DBSCAN and hierarchical clustering methods. **A.1.** In the given example, with four-period repetition, the question of finding two classes within one period becomes the question of finding $10 = 2 \times (4+1)$ classes within the total $5 = 4+1$ periods with the K-means method. Points with the same color belong to the same class. We can locate the correct cluster output in the middle (i.e., 3rd) period. **A.2.** We apply the same DBSCAN parameters as Fig. 1 ($\epsilon = 0.5$ and $n_{\min} = 2$) on the rectangular plane with five periods. Points with the same color belong to the same class. We consistently identify the same pattern for class 1 (four points are classified into the same class repetitively five times, highlighted as blue shadow). All other points that are unclassified are defined as outliers, which are also the outliers defined by given parameters (referred to as label '-1'). **A.3.** We apply DBSCAN with different parameters ($\epsilon = 1.5$ and $n_{\min} = 2$) on the rectangular plane with five periods. Two patterns have been identified this time and those two patterns include all 7 sample points, so two classes are found with no outliers. **B.1.** Clustering result of K-means in **A.1**. **B.2.** Clustering result of DBSCAN in **A.2**. **B.3.** Clustering result of DBSCAN in **A.3**. **B.4.** Clustering result of the hierarchical clustering method with the proposed algorithm, its result is improved from the result shown in Fig. 1**C.3**, and its result is consistent with the K-means and DBSCAN methods where both class 1 and class 2 are included as sub-class (highlighted in red).

Algorithm 2 Search algorithm for DBSCAN

Require: $c, c \in \mathbb{Z}^+$, clustering result of $c+1$ periods DB_t
Ensure: Final clustering DB_c
number of class has been identified: $k \leftarrow \text{unique}(\text{DB}_t)$
samples points within each class P_k
map all points in P_k to the original period $[0°, 360°)$
count the class has the same points Z_k
setup the repetition threshold Y ▷ based on c
for $x = 1$ to k **do**
  if $Z_x \geq Y$ **then**
    class x should be included in DB_c
  end if
end for
points not included in DB_c are outliers

Algorithm 3 Search algorithm for Hierarchical Clustering

Require: $c, c \in \mathbb{Z}^+$, all points P within $c+1$ period
Ensure: Final dendrogram HC_c
calculate pairwise distance by reconstructed coordinates: $Y \leftarrow \text{pdist}(P)$
pairwise combination within all sample points P: $C \leftarrow$ nchoosek(1:size(P),2)
pairwise combination within the original sample points P_1: $C_1 \leftarrow$ nchoosek(1:size(P_1),2)
map pairwise combination P to the original period: $P' \leftarrow \mod(P, \text{size}(P_1))$
Find the minimal pairwise distance calculated
for $x = 1$ to size(C_1) **do**
  $Y_{C_1} \leftarrow \min(Y_x)$ ▷ pairwise distance of C_1
end for
build the dendrogram by Y_{C_1}: $Z \leftarrow \text{linkage}(Y_{C_1})$

3.4 Validation Metrics

The Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) are both external validation metrics used to assess the quality of clustering results by comparing them to ground truth or benchmark labels. The ARI is a commonly employed metric in cluster validation. It serves as a valuable tool for identifying potential issues within classification algorithms, particularly when utilized in conjunction with conventional performance metrics [35]. The NMI assesses the mutual information existing between the authentic labels and the labels ascribed by a clustering algorithm, while simultaneously accounting for label distribution disparities. It quantifies how much information is shared between the two sets of labels [36]. Both ARI and NMI yield values within the range of 0 to 1. An ARI value of 1 signifies perfect congruence between a partition and the intrinsic structure, while values close to 0 indicate a random partition (in practice, a negative value means it is a worse choice than the expected value of random labels [37]). Similarly, NMI assumes a value of 0 when there is no mutual information (corresponding to random labeling) and attains a value of 1 in scenarios where the clustering aligns flawlessly with the ground truth.

4 Results

In the previous section, we have shown that the proposed method provides the correct classification for the given example with the K-means, DBSCAN, and hierarchical clustering methods (see Fig. 4). Next, we provide additional evidence of the value of the approach on both synthetic and real data.

4.1 Application to Synthetic Data

The previous example only included seven points to illustrate the problem and introduce the framework. Now, we examine the clustering performance under different distribution scenarios.

4.1.1 Larger sample size

Fifty points were randomly generated on the unit circle ($r = 1$) based on two classes. In Fig. 5**A**, we show how different R (introduced in (5)) would shape the clustering output with the K-means method. Smaller R gives more balanced weight between r and θ, while larger R favors angle-driven (emphasize on differentiating θ) clustering results. The flexibility in the approach enables one to incorporate prior knowledge about the relative importance of both dimensions. The angle-driven approach (greater R) produces result identical to conventional K-means angle-focused circular clustering algorithms (e.g., Fast Optimal Circular Clustering (FOCC), brute force optimal circular clustering (BOCC), and heuristic circular clustering (HEUC) [14], see TABLE 1 and Fig. 1**A.1** in Appendix A). In addition, a balanced weight for R generates the same two classes from the underlying generative model, aligning with ground truth (see TABLE 2).

4.1.2 Multiple classes ($K \geq 3$)

During the previous discussion, only considered only clustering problems with two default classes. Here we use our proposed method on a dataset generated based on five classes. More period repetitions are needed to cluster a larger number of classes. Simulation results show that both K-means and DBSCAN uncover the same clusters as the generating distribution (see Fig. 5**B**). Additionally, for hierarchical clustering, the clades provide the same information as K-means and DBSCAN, where five clades are formed that include ten points within each class (see TABLE 1, TABLE 2, and Appendix A).

4.2 Application to Neural Data

In a previous clinical study, we measured the inter-trial phase coherence (ITPC) of the quasi-alpha oscillation (6-13Hz) from participants' EEG signals [9, 22, 38]. In this study, we hypothesized that stronger and more precise inter-trial coherence would be a biomarker of the efficacy of a neurostimulation therapeutic. By mapping the participant's post-stimulation quasi-alpha phase onto the unit circle and calculating the circular mean, we obtain 30 points (i.e., 30 experimental sessions) in polar coordinates. In this representation a larger $r \in [0, 1]$ indicates consistency of each session's phase synchronization across all trials (greater r means better synchronization) and θ refers to the phase of

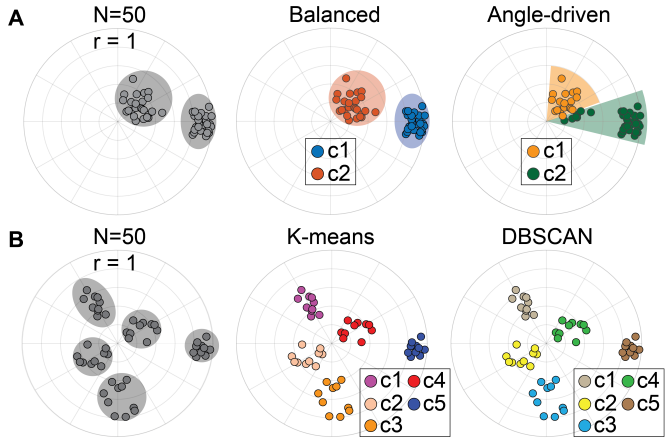

Fig. 5. Application to synthetic data ($N = 50$). **A.** Clustering results with K-means on two groups. Firstly, 50 points are randomly generated based on two classes on the unit cycle (highlighted with the gray shadow). With a smaller R applied, the clustering results have a balanced weight between r and θ, while angle-driven clustering results will be generated when a larger R is used. The angle-driven clustering result is identical to the result generated by FOCC, BOCC, and HEUC [14]. **B.** Clustering results of multiple groups with K-means and DBSCAN. Another 50 points are randomly generated based on five classes (highlighted with the gray shadow) on the unit cycle. Applying both K-means and DBSCAN on our reconstructed representation yields the ground truth as output. More quantitative comparison results can be found at TABLE 1 and TABLE 2.

TABLE 1
The Comparison of Clustering Results on Synthetic Data (angle-driven)

	Validation Metrics			
	$\mathbf{ARI}^{(1)}$	$\mathbf{NMI}^{(1)}$	$\mathbf{ARI}^{(2)}$	$\mathbf{NMI}^{(2)}$
K-means	0.99±0.02	0.99±0.04	0.98±0.06	0.93±0.12
FOCC	0.99±0.02	0.99±0.04	0.98±0.04	0.94±0.10
BOCC	0.99±0.02	0.99±0.04	0.98±0.06	0.93±0.11
HEUC	0.99±0.02	0.99±0.04	0.98±0.06	0.92±0.16

Ten random seeds are used to generate the synthetic data. The same parameters are used for clustering each time (e.g., parameter R and the number of period repetitions). $^{(1)}$Fifty points with different angles are randomly generated based on two classes within the unit cycle. $^{(2)}$Fifty points with different angles but the same radius are randomly generated based on 5 classes within the unit cycle. Both Indices show strong agreement between clustering and ground truth. No significant difference was observed among methods.

the synchronization (more details are available in [38]). We applied our proposed method to these data to assess the phase synchronization among all experimental sessions for each subject.

As shown in Fig. 6, when assuming the number of classes is predetermined ($K = 2$, inlier and outlier), the angle-focused K-means method yields the same clustering result as the FOCC method [14] for both subjects. Moreover, the representation can also be used with DBSCAN when density-based methods are desirable (see Fig. 6). Interestingly, the clustering result of Sub#1 shows consistent similarity within blue points (similar class 'c1' is defined with angle-focused K-means, DBSCAN(1), DBSCAN(2), and balanced K-means in Appendix A), which is consistent with our hypothesis and clinical experimental results–i.e. Sub#1 is a subject who showed greater phase synchronization and better clinical improvement from the neurostimulation treatment.

TABLE 2
The Comparison of Clustering Results on Synthetic Data (balanced)

	Validation Metrics			
	$\mathbf{ARI}^{(1)}$	$\mathbf{NMI}^{(1)}$	$\mathbf{ARI}^{(2)}$	$\mathbf{NMI}^{(2)}$
K-means^a	0.99±0.02	0.99±0.04	0.90±0.12	0.90±0.10
DBSCAN^a	0.98±0.05	0.98±0.06	0.90±0.12	0.93±0.09
Hierarchical^a	0.98±0.03	0.98±0.05	0.89±0.12	0.91±0.11
K-means^b	0.01±0.08	0.01±0.02	-0.01±0.02	0.01±0.00
DBSCAN^b	0.02±0.05	0.00±0.02	-0.01±0.01	0.00±0.01
Hierarchical^b	0.01±0.08	0.01±0.02	-0.01±0.03	0.00±0.01

Ten random seeds are used to generate the synthetic data. The same parameters are used for clustering each time (e.g., parameter R and the number of period repetitions). $^{(1)}$Fifty points with different angles and radius are randomly generated based on two classes within the unit cycle. $^{(2)}$Fifty points with different angles and radii are randomly generated based on 5 classes within the unit cycle. Both Indices show strong agreement between clustering and ground truth. No significant difference was observed among methods. ^{a}Clustering method applied with our proposed framework. ^{b}Clustering method directly applied to the original data. Our framework is significantly better (at the confidence level of 99.99%) than the direct application.

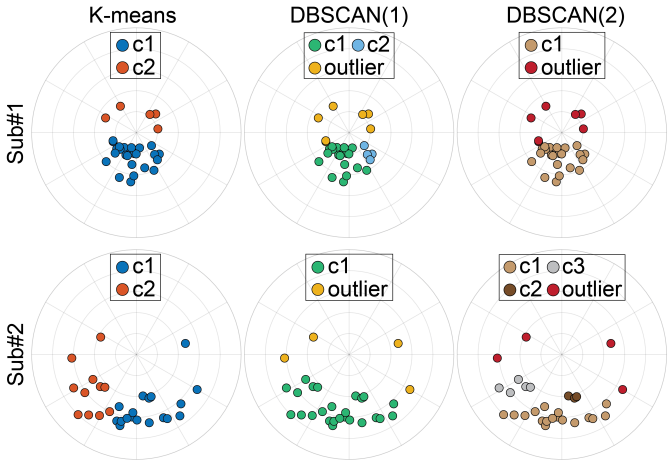

Fig. 6. Application to neural data ($N = 30$). Two subjects are used as examples. The output of K-means with a larger R to capture the difference in θ is identical to the output of another angle-focused circular clustering method (FOCC with $K = 2$, see Fig. 1**B.1** in Appendix A). Additionally, DBSCAN with two sets of parameters is examined. DBSCAN(1) is generated with $\epsilon = 0.25$ and $n_{\min} = 5$, and DBSCAN(2) is generated with $\epsilon = 0.2$ and $n_{\min} = 3$ for both subjects. The DBSCAN clustering result of Sub#1 shows this subject has consistent similarity within c1 class (green points in DBSCAN(1) and camel points in DBSCAN(2)), while no such observation is seen in Sub#2.

4.3 Application to DNA Data

Another common application of circular clustering in biology is the analysis of DNA or RNA sequences. The circular representation of DNA or RNA sequences allows for the visualization of the periodicity or cyclic patterns that may be present. Clustering techniques can be applied to these circular sequences to identify similar patterns or motifs, which can provide insights into the structure and function of the sequences [10, 39]. With the polar coordinate representation of DNA sequence proposed by Dai et al. [40] (or other methods to construct circular structure [10, 11]), we can map any given biological sequence into polar coordinates where the radius and angles are determined by the distribution of the dual nucleotides (e.g., 'AC', 'AG', etc). Then by

applying hierarchical clustering, we can gain insight into the relationships between different sequences and patterns in the sequence distribution.

Fig. 7 presents the results of hierarchical clustering analysis on the first exon of β-globin gene of Human, Chimpanzee, and Mouse [41]. To compare the similarity between the dendrograms obtained for the three species, we computed the cophenetic correlation matrix [42]. The result revealed that the dendrograms for Chimpanzee and Mouse are both similar to Human and there is a higher degree of similarity between the dendrogram for Human and Chimpanzee (cophenetic correlation: 0.944) than Human and Mouse (cophenetic correlation: 0.927, see Fig. 2 in Appendix B). These findings were consistent with the results obtained using other metrics [40, 41].

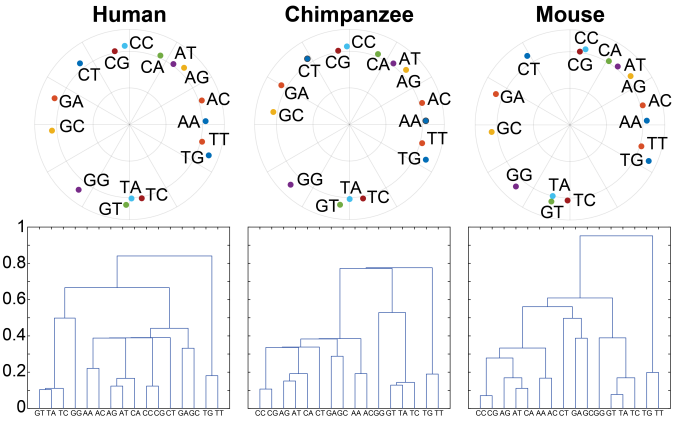

Fig. 7. Hierarchical clustering result on the coding sequences of the first exon of β-globin gene of Human, Chimpanzee, and Mouse. The top raw shows the polar coordinates representation of the dual nucleotides. The bottom row is the corresponding dendrogram of three species.

5 Discussion

Analysis of circular data is of increasing interest in multiple field in the life sciences. Due to the nonequivalent representation caused by converting polar coordinates to Cartesian coordinates (e.g., $\pi/2 \neq 3\pi/2$ but $\cos(\pi/2) = \cos(3\pi/2)$), most traditional clustering algorithms fail if the difference in angle θ plays an important role. Several circular clustering algorithms have been proposed to address this issue, but they also have disadvantages, such as being angle-focused or lacking generality [2, 3, 14, 29, 30]. Our proposed method addresses these problems by adapting the cylindrical coordinate system [43, 44] and taking advantage of fundamental mathematical properties of polar coordinates that have been observed and successfully applied in phase-controlled coordinate choice optimization [34, 45].

Our method marks a departure from conventional methods in several fundamental ways. Firstly, unlike most existing techniques that attempt to address circular data within a Cartesian coordinate system [24], our method is directly designed based on polar coordinates. This direct consideration of polar coordinates offers a more natural and efficient approach for handling circular data patterns, eliminating the need for complex and computationally expensive distance metric modifications [3, 30]. Furthermore, our approach stands out by leveraging the data representation itself to account for circularity. Instead of merely focusing on updating distance metrics, we fundamentally simplify the problem by using the inherent mathematical property of the data. The proposed framework does not make any assumptions about the distance metric selection and it completely depends on the applied clustering algorithms (during testing, the default distance measure we used is the Euclidean distance, which is the most common distance metric). This innovative strategy not only streamlines the clustering process but also ensures greater flexibility and adaptability when applying various clustering methods. This adaptability ensures that our approach can cater to diverse research needs and accommodate various data types and structures. Lastly, our method offers flexibility in parameter selection. Researchers can dynamically adjust the parameter R based on the benchmark clustering outcomes, literature evidence, or their desired classification outcomes in practice, enhancing accuracy and providing insights into the biological interpretations of the data. This unique feature empowers researchers to tailor their analyses to suit specific research objectives, ultimately advancing the understanding of complex data patterns.

In the proposed coordinate reconstruction, varying the intermediary size of the cylindrical representation (through parameter R) helps to assign different weights to θ based on the research problem. With enough period repetitions, we can always find the accurate clustering result within the given dataset. Furthermore, our method is able to handle large sample sizes and multiple target classes, simply by applying a greater number of period repetitions. The computational complexity of the proposed method is relatively low, being linear in time as one increases the repetition number. The reconstruction is designed to be general for a polar coordinate system without any embedding context, so it can be broadly applied to circular clustering problems on different representations and in different fields [14, 16, 17, 29, 46].

6 Conclusion

In this paper, we propose a technique to reconstruct polar coordinates with appropriate period repetitions to form a better representation for more accurate circular clustering. This novel tool for the analysis of circular data is useful, accurate, and generally applicable. Most importantly, it can be easily interpreted and modified for various analysis needs.

Software Availability

All presented and evaluated algorithms are implemented in Matlab2022b (Github). Fig. 1**A.1**, Fig. 1**B.1**, and Fig. 2 in Appendix are visualized with R programming languages (package 'OptCirClust': https://cran.r-project.org/web/packages/OptCirClust/index.html and 'dendrogram': https://CRAN.R-project.org/package=dendextend). The 'sklearn.metrics' package is employed (https://scikit-learn.org/stable/about.html#citing-scikit-learn) in Python 3.6 for the calculation of Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI).

ACKNOWLEDGMENTS

This work was funded by the National Institute of Mental Health (MH106775), a Vannevar Bush Faculty Fellowship from the US Department of Defense (N00014-20-1-2027), and a Center of Excellence grant from the Air Force Office of Scientific Research (FA9550-22-1-0337). We would like to thank Sharath Koorathota for providing feedback on the manuscript draft. We would like to thank our collaborators at the Medical University of South Carolina for their help with the data collection of our phase synchronization study.

REFERENCES

[1] S. Anand, P. Padmanabham, and A. Govardhan, "Effect of distance measures on partitional clustering algorithms using transportation data," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 6, pp. 5308–5312, 2015.

[2] R. Merrell and D. Diaz, "Clustering analyses methods: strategies and algorithms," *Reviews in Theoretical Science*, vol. 4, no. 2, pp. 153–158, 2016.

[3] Y. S. Patil and M. R. Joshi, "Clustering with polar coordinates system: Exploring possibilities," in *Smart Intelligent Computing and Applications*, pp. 553–560, Springer, 2019.

[4] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[5] J. Fell and N. Axmacher, "The role of phase synchronization in memory processes," *Nature reviews neuroscience*, vol. 12, no. 2, pp. 105–118, 2011.

[6] C. C. Canavier, "Phase-resetting as a tool of information transmission," *Current opinion in neurobiology*, vol. 31, pp. 206–213, 2015.

[7] F. Fröhlich and D. A. McCormick, "Endogenous electric fields may guide neocortical network activity," *Neuron*, vol. 67, no. 1, pp. 129–143, 2010.

[8] S. E. Qasim, I. Fried, and J. Jacobs, "Phase precession in the human hippocampus and entorhinal cortex," *Cell*, vol. 184, no. 12, pp. 3242–3255, 2021.

[9] S. P. Pantazatos, J. R. McIntosh, G. T. Saber, X. Sun, J. Doose, J. Faller, Y. Lin, J. B. Teves, A. Blankenship, S. Huffman, *et al.*, "Functional and effective connectivity between dorsolateral prefrontal and subgenual anterior cingulate cortex depends on the timing of transcranial magnetic stimulation relative to the phase of prefrontal alpha eeg," *bioRxiv*, 2022.

[10] L. S. Kristensen, M. S. Andersen, L. V. Stagsted, K. K. Ebbesen, T. B. Hansen, and J. Kjems, "The biogenesis, biology and characterization of circular rnas," *Nature Reviews Genetics*, vol. 20, no. 11, pp. 675–691, 2019.

[11] H. Kim, N.-P. Nguyen, K. Turner, S. Wu, A. D. Gujar, J. Luebeck, J. Liu, V. Deshpande, U. Rajkumar, S. Namburi, *et al.*, "Extrachromosomal dna is associated with oncogene amplification and poor outcome across multiple cancers," *Nature genetics*, vol. 52, no. 9, pp. 891–897, 2020.

[12] R. Dong, L. He, R. L. He, and S. S.-T. Yau, "A novel approach to clustering genome sequences using internucleotide covariance," *Frontiers in Genetics*, vol. 10, p. 234, 2019.

[13] K. W. Govek, V. S. Yamajala, and P. G. Camara, "Clustering-independent analysis of genomic data using spectral simplicial theory," *PLoS computational biology*, vol. 15, no. 11, p. e1007509, 2019.

[14] T. Debnath and M. Song, "Fast optimal circular clustering and applications on round genomes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 6, pp. 2061–2071, 2021.

[15] K. Sivasankar, M. Sujaritha, P. Pasupathi, and S. Muthukumar, "Retraction notice: Fcm based iris image analysis for tissue imbalance stage identification," in *2012 International Conference on Emerging Trends in Science, Engineering and Technology (INCOSET)*, pp. 1–1, IEEE, 2012.

[16] S. E. Hussein, O. A. Hassan, and M. H. Granat, "Assessment of the potential iridology for diagnosing kidney disease using wavelet analysis and neural networks," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 534–541, 2013.

[17] T. Deya, S. Biswas, S. Chatterjee, S. Manna, A. Nandye, and S. C. Basake, "2d polar co-ordinate representation of amino acid sequences with some applications to ebola virus, sars and sars-cov-2 (covid-19)," in *MOL2NET, Int. Conf. Multidisciplinary Sci*, 2020.

[18] S. C. Basak, T. Dey, A. Nandy, *et al.*, "Cluster analysis of coronavirus sequences using computational sequence descriptors: With applications to sars, mers and sars-cov-2 (covid-19)," *Current Computer-Aided Drug Design*, vol. 17, no. 7, pp. 936–945, 2021.

[19] A. Nandy, "Mapping biomolecular sequences: Graphical representations-their origins, applications and future prospects," *Combinatorial Chemistry & High Throughput Screening*, vol. 25, no. 3, pp. 354–364, 2022.

[20] L. Berlincioni, F. Becattini, L. Seidenari, and A. Del Bimbo, "Multiple future prediction leveraging synthetic trajectories," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 6081–6088, IEEE, 2021.

[21] M. S. George, S. Huffman, J. Doose, X. Sun, M. Dancy, J. Faller, X. Li, H. Yuan, R. Goldman, P. Sajda, *et al.*, "Eeg synchronized left prefrontal transcranial magnetic stimulation (tms) for treatment resistant depression is feasible and produces an entrainment dependent clinical response: A randomized controlled double blind clinical trial," *Available at SSRN 4334289*, 2023.

[22] X. Sun, J. Doose, J. Faller, J. McIntosh, G. T. Saber, S. Huffman, S. Pantazatos, H. Yuan, R. Goldman, T. Brown, *et al.*, "Increased entrainment and decreased excitability predict efficacious treatment of closed-loop phase-locked rtms for treatment-resistant depression," *medRxiv*, pp. 2023–10, 2023.

[23] K. V. Mardia, P. E. Jupp, and K. Mardia, *Directional statistics*, vol. 2. Wiley Online Library, 2000.

[24] D. Charalampidis, "A modified k-means algorithm for circular invariant clustering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 12, pp. 1856–1865, 2005.

[25] G. Armano and M. R. Farmani, "Multiobjective clustering analysis using particle swarm optimization," *Expert Systems with Applications*, vol. 55, pp. 184–193, 2016.

[26] X. Xu, J. Li, M. Zhou, J. Xu, and J. Cao, "Accelerated

two-stage particle swarm optimization for clustering not-well-separated data," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 11, pp. 4212–4223, 2018.

[27] S.-J. Chang-Chien, W.-L. Hung, and M.-S. Yang, "On mean shift-based clustering for circular data," *Soft Computing*, vol. 16, pp. 1043–1060, 2012.

[28] C. Abraham, N. Molinari, and R. Servien, "Unsupervised clustering of multivariate circular data," *Statistics in medicine*, vol. 32, no. 8, pp. 1376–1382, 2013.

[29] M. Li, M. Stolz, Z. Feng, M. Kunert, R. Henze, and F. Küçükay, "An adaptive 3d grid-based clustering algorithm for automotive high resolution radar sensor," in *2018 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, pp. 1–7, IEEE, 2018.

[30] F. Wang, Y. Xie, Z. Hu, K. Zhang, and Y. Zhang, "An adaptive clustering algorithm based on circular units," in *2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pp. 179–186, IEEE, 2021.

[31] M. Ghahramani, A. O'Hagan, M. Zhou, and J. Sweeney, "Intelligent geodemographic clustering based on neural network and particle swarm optimization," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 6, pp. 3746–3756, 2021.

[32] J. Liu, X. Liu, S. Wang, S. Zhou, and Y. Yang, "Hierarchical multiple kernel clustering," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, pp. 8671–8679, 2021.

[33] X. Wu, Z. Ren, and F. R. Yu, "Parameter-free shifted laplacian reconstruction for multiple kernel clustering," *IEEE/CAA Journal of Automatica Sinica*, 2023.

[34] B. Lin, B. Chong, Y. Ozkan-Aydin, E. Aydin, H. Choset, D. I. Goldman, and G. Blekherman, "Optimizing coordinate choice for locomotion systems with toroidal shape spaces," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7501–7506, IEEE, 2020.

[35] J. M. Santos and M. Embrechts, "On the use of the adjusted rand index as a metric for evaluating supervised classification," in *International conference on artificial neural networks*, pp. 175–184, Springer, 2009.

[36] Z. F. Knops, J. A. Maintz, M. A. Viergever, and J. P. Pluim, "Normalized mutual information based registration using k-means clustering and shading correction," *Medical image analysis*, vol. 10, no. 3, pp. 432–439, 2006.

[37] J. E. Chacón and A. I. Rastrojo, "Minimum adjusted rand index for two clusterings of a given size," *Advances in Data Analysis and Classification*, vol. 17, no. 1, pp. 125–133, 2023.

[38] J. Faller, J. Doose, X. Sun, J. R. Mclntosh, G. T. Saber, Y. Lin, J. B. Teves, A. Blankenship, S. Huffman, R. I. Goldman, *et al.*, "Daily prefrontal closed-loop repetitive transcranial magnetic stimulation (rtms) produces progressive eeg quasi-alpha phase entrainment in depressed adults," *Brain Stimulation*, vol. 15, no. 2, pp. 458–471, 2022.

[39] Z.-H. Qi, J.-M. Wang, and X.-Q. Qi, "Classification analysis of dual nucleotides using dimension reduction," *Journal of theoretical biology*, vol. 260, no. 1, pp. 104–109, 2009.

[40] Q. Dai, X. Guo, and L. Li, "Sequence comparison via polar coordinates representation and curve tree," *Journal of theoretical biology*, vol. 292, pp. 78–85, 2012.

[41] N. Jafarzadeh and A. Iranmanesh, "A new measure for pairwise comparison of protein sequences," *MATCH: Communications in Mathematical and in Computer Chemistry*, vol. 74, pp. 563–574, 2015.

[42] R. R. Sokal and F. J. Rohlf, "The comparison of dendrograms by objective methods," *Taxon*, pp. 33–40, 1962.

[43] A. Roy, S. K. Parui, and U. Roy, "Swgmm: a semi-wrapped gaussian mixture model for clustering of circular–linear data," *Pattern Analysis and Applications*, vol. 19, pp. 631–645, 2016.

[44] A. Roy, A. Pal, and U. Garain, "Jclmm: A finite mixture model for clustering of circular-linear data and its application to psoriatic plaque segmentation," *Pattern recognition*, vol. 66, pp. 160–173, 2017.

[45] B. Chong, Y. O. Aydin, C. Gong, G. Sartoretti, Y. Wu, J. M. Rieser, H. Xing, J. W. Rankin, K. Michel, A. G. Nicieza, *et al.*, "Coordination of back bending and leg movements for quadrupedal locomotion.," in *Robotics: Science and Systems*, vol. 20, Pittsburgh, PA, 2018.

[46] F. Rijo-Ferreira and J. S. Takahashi, "Genomics of circadian rhythms in health and disease," *Genome medicine*, vol. 11, pp. 1–16, 2019.

Xiaoxiao Sun received the bachelor's degree in Mathematics and Economics from Baruch College, City University of New York, and a master's degree in Biomedical Engineering from Columbia University, in 2018 and 2021, respectively. Currently, she is working toward the PhD degree in Biomedical Engineering with Columbia University. She is interested in biomedical imaging, bioinformatics, neural engineering, neuromodulation and brain-computer interface (BCI). Much of her research is focused on using neuroimaging (such as EEG and fMRI) as well as other behavioral and physiological measures (e.g., eye-tracking) to understand the brain dynamics at rest and during decision-making. Her recent work is about integrating neurostimulation, such as transcranial magnetic stimulation (TMS), as a potential tool for inferring causal relationships in these neural circuits so as to improve the efficacy of neurostimulation for the treatment of psychiatric diseases, such as major depressive disorder (MDD).

Paul Sajda is the Vikram S. Pandit Professor of Biomedical Engineering and Professor of Electrical Engineering and Radiology (Physics) at Columbia University. He received a BS in electrical engineering from MIT in 1989 and an MSE and Ph.D. in bioengineering from the University of Pennsylvania in 1992 and 1994, respectively. Professor Sajda is interested in what happens in our brains when we make a rapid decision and what neural processes and representations drive our underlying preferences and choices, mainly under time pressure. His work in understanding the basic principles of rapid decision-making in the human brain relies on measuring human subject behavior simultaneously with cognitive and physiological states. Professor Sajda co-founded several neurotechnology companies and works closely with various scientists and engineers, including neuroscientists, psychologists, computer scientists, and clinicians. He is a fellow of the IEEE, AMBIE, and AAAS. He also received the Vannevar Bush Faculty Fellowship (VBFF), the DoD's most prestigious single-investigator award. Professor Sajda is also the current President of IEEE EMBS.