

EXPLORATORY DATA ANALYSIS WITH APPLICATIONS TO BIOINFORMATICS

KANTI V. MARDIA, JOHN T. KENT, ZHENGZHENG ZHANG
AND CHARLES C. TAYLOR
DEPARTMENT OF STATISTICS, UNIVERSITY OF LEEDS, LEEDS, UK.

ABSTRACT. There are many statistical challenging problems in protein and RNA Bioinformatics. We can regard the backbone of a typical protein as an articulated object in three dimensions with fixed bond lengths between successive amino acids. Hence it can be viewed as a long time series (with hundreds or thousands of amino acids), where all the information lies in the angles between successive bonds. There are two types of angles: bond or planar angles, analogous to colatitude, which are nearly constant here; and dihedral angles, analogous to longitude, which contain all the information. Thus the basic protein description is reduced to a circular time series. There is also further angular information coming from side chains.

The aim is to find patterns in such data (see for example, Boomsma et al, 2008). This report will give new methods and visualisation tools, including time series analysis, circular principal component analysis and clustering (cf: Mu et al, 2005; Altis et al, 2007), together with examples.

Keywords: Circular time series analysis, circular principal component analysis, Torus distance and clustering analysis.

1. INTRODUCTION

The first part (Sections 2-4) considers the problem of measuring the association between several variables in a vector-valued stochastic processes. The basic RNA description is a time series with blocks of 6 different types of dihedral angles from the backbone. These correlation measures are then applied to the RNA data set obtained from Frellsen *et al.* (2009). Our objective is to investigate the relationship between different types of dihedral angles.

The second part (Sections 5-8) is concerned with principal component analysis on angular data. We examine the existing circular PCA methods and its properties. We then apply these method on a variety of the data sets simulated from both the sine and cosine distribution. Finally, we discuss the strength and weakness of each method.

The third part (Sections 9-11) is concerned with cluster analysis on angular data. We define the “similarity” between any pair of data points on the p -fold torus by calculating the angular distance between them in Section 9. After that, we give a brief review about the hard clustering algorithm in Section 10. Finally, we apply the partitioning

around medoids algorithm to the RNA and protein side-chain data sets, based on the angular similarity measure.

2. A CIRCULAR CORRELATION COEFFICIENT

Let X, Y be two random variables from a bivariate distribution, $F(X, Y)$, on the torus, where $-\pi < X, Y \leq \pi$. μ, ν are the circular mean directions of the X, Y , respectively. The section 8.2 of Jammalamadaka & Sarma (1988) gives a circular correlation coefficient:

$$\rho_{X,Y} = \frac{E(\sin(X - \mu) \sin(Y - \nu))}{\sqrt{\text{Var}(\sin(X - \mu)) \text{Var}(\sin(Y - \nu))}}. \quad (1)$$

Note that this circular correlation coefficient is naturally analogous to Pearson's correlation coefficient in the linear case, since $\sin(X - \mu)$ and $\sin(Y - \nu)$ measure the standard deviations of X and Y from the circular mean directions μ and ν . Further,

$$E(\sin(X - \mu)) = E(\sin(Y - \nu)) = 0$$

which is analogous to the fact that the first central moments in the linear case are 0. Note that Mardia & Jupp (1999) also discussed other measures to compute circular-circular correlation coefficients such as the correlation coefficient based on embedding approach and the rank correlation coefficient.

The circular correlation coefficient is re-written in Jammalamadaka *et al.* (2001) as

$$\rho_{X,Y} = \frac{E[\cos(X - Y - \mu + \nu) - \cos(X + Y - \mu - \nu)]}{2\sqrt{E(\sin^2(X - \mu))E(\sin^2(Y - \nu))}}. \quad (2)$$

Since $E(\cos(X - \mu))$ is a measure of concentration of X around the mean direction, the first term in the numerator measures how strongly the distribution of $(X - \mu) - (Y - \nu)$ is concentrated and this contributes to the positive correlation. Similarly the second term measures the negative part of the correlation.

The circular correlation coefficient has the following properties given in Jammalamadaka *et al.* (2001):

- (1) $\rho_{X,Y}$ does not depend on the zero direction used for either variable;
- (2) $\rho_{X,Y} = \rho_{Y,X}$;
- (3) $|\rho_{X,Y}| \leq 1$;
- (4) $\rho_{X,Y} = 0$ if X and Y are independent although the converse need not be true (this is because the sine function is not bijective);
- (5) If X and Y have full support, $\rho_{X,Y} = 1$ iff $Y = X + \text{const} \pmod{2\pi}$ and $\rho_{X,Y} = -1$ iff $X + Y = \text{const} \pmod{2\pi}$.

Assume that a random sample $(x, y) = \{(x_i, y_i) : i = 1, 2, \dots, n\}$ is drawn from a bivariate distribution, $F(X, Y)$, on the torus. Then, the sample circular correlation coefficient is given by

$$r_{x,y} = \frac{\sum_{i=1}^n \sin(x_i - \bar{x}) \sin(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n \sin^2(x_i - \bar{x}) \sin^2(y_i - \bar{y})}} \quad (3)$$

where \bar{x} and \bar{y} are the (circular) sample mean directions of x and y . Moreover, Jam-malamadaka *et al.* (2001) developed a significance test for the sample measure, $r_{x,y}$. Their Corollary 8.1 stated that

Corollary 1. Under the hypothesis $H_0 : \rho_{x,y} = 0$ for large n ,

$$\sqrt{nr_{x,y}} \sim N\left(0, \frac{\lambda_{22}}{\lambda_{20}\lambda_{02}}\right) \quad (4)$$

and by Slutsky's theorem,

$$\sqrt{n} \sqrt{\frac{\hat{\lambda}_{20}\hat{\lambda}_{02}}{\hat{\lambda}_{22}}} r_{x,y} \sim N(0, 1), \quad (5)$$

where

$$\begin{aligned} \lambda_{20} &= E\{\sin^2(X - \mu)\}, & \lambda_{02} &= E\{\sin^2(Y - \nu)\}, & \lambda_{22} &= E\{\sin^2(X - \mu) \sin^2(Y - \nu)\} \\ \hat{\lambda}_{20} &= \frac{1}{n} \sum_{k=1}^n \sin^2(x_k - \bar{x}), & \hat{\lambda}_{02} &= \frac{1}{n} \sum_{k=1}^n \sin^2(y_k - \bar{y}), & \hat{\lambda}_{22} &= \frac{1}{n} \sum_{k=1}^n \sin^2(x_k - \bar{x}) \sin^2(y_k - \bar{y}). \end{aligned}$$

In particular, if X, Y are two variables of the bivariate sine distribution, for high concentration, we have

$$E(\sin^2 X) \cong E(X^2); \quad E(\sin^2 Y) \cong E(Y^2).$$

For $\rho = 0$ (i.e., $\lambda = 0$), X and Y are independent, and we then have

$$\lambda_{22} = E(\sin^2 X \sin^2 Y) = E(\sin^2 X)E(\sin^2 Y).$$

Therefore, $\lambda_{22}/(\lambda_{02}\lambda_{20}) = 1$, so $\sqrt{nr_{x,y}} \sim N(0, 1)$ for large n .

This circular correlation coefficient $r_{x,y}$ and its significance test are implemented in the R package of CircStats.

For small x, y , $\sin(x), \sin(y)$ are approximately equal to x, y . This means that the sample circular correlation coefficient of x, y , $r_{x,y}$, is approximately to Pearson's sample correlation coefficient, if x, y are all small. However, if x goes from $\pi/2$ to π , then $\sin x$ decreases from 1 to 0 reversely. Clearly, sine function is not bijective, i.e., the two different values of x may map to the same value of $\sin(x)$. It may worth to explore the circular correlation coefficient of x, y when the data are not concentrated.

Let us consider the following example. Consider a bivariate cosine distribution (Mardia *et al.*, 2007) with density given as:

$$f_c(\theta_1, \theta_2) = \tilde{C}_2^{-1} \exp\{\kappa_1 \cos(\theta_1 - \mu_1) + \kappa_2 \cos(\theta_2 - \mu_2) - \kappa_3 \cos(\theta_1 - \mu_1 - \theta_2 + \mu_2)\}, \quad (6)$$

where $\kappa_1, \kappa_2 > 0$, $-\pi < \mu_1, \mu_2 \leq \pi$ and \tilde{C}_2^{-1} is the normalizing constant that does not have a closed form. Let

$$\kappa_1 = 0.5, \quad \kappa_2 = 0.5, \quad \kappa_3 = 25.$$

That is, κ_1 and κ_2 are small but κ_3 is comparatively large. Note that it is always possible to move the sample mean directions to the origin. Without loss of generality, we assume the circular mean directions $\mu_1 = \mu_2 = 0$. A random sample of 1000

data pairs, (x, y) , is then simulated from the bivariate cosine distribution using the acceptance-rejection method (see web appendix of Mardia *et al.* (2007) for details).

The pairwise scatter plot of the sample is shown on the left of Figure 1, whereas the sine transformation is applied to the sample, and produces the scatter plot on the right of the figure. We can see from the left plot that the data is semi-diffused, i.e., the data is concentrated in one direction, but is diffused in the another direction perpendicular to the previous one. There are artificial cuts at $-\pi$ and π , so these data points are discontinuous at the boundaries splitting into two data clouds. In fact, the data points will form a circle on a 3D torus. The transformed data points, on the right of the figure, form an elliptic ring, at which we observe high concentration at $(-1, -1)$ and $(1, 1)$. This results from the sine transformation. Suppose we have a random sample

$$\alpha \sim \text{Uniform}(-\pi, \pi),$$

then $\sin(\alpha)$ will be concentrated around -1 and 1 when $\alpha = -\pi/2$ or $\pi/2$. After that, the circular correlation coefficient is calculated using (3), and we have $r_{x,y} = 0.9380$ that is consistent with what we observe from the left plot of the figure, i.e., there exists a strong positive correlation between x and y . Finally, we perform a significant test for the sample measure, $r_{x,y}$. Under the hypothesis $H_0 : \rho_{x,y} = 0$, we have a test statistic $\sqrt{n} \sqrt{\frac{\hat{\lambda}_{20}\hat{\lambda}_{02}}{\hat{\lambda}_{22}}} r_{x,y} = 33.80$, which is tested against the 5% critical value 1.96 provided by the standard normal distribution. The corresponding p -value is close to 0, so we are going to reject the null hypothesis that $\rho_{x,y} = 0$.

3. CIRCULAR CORRELATION MATRIX FUNCTIONS

Let $\mathbf{Z}_t = [Z_{1,t}, Z_{2,t}, \dots, Z_{m,t}]$ be an m -dimensional jointly stationary real-valued vector process with mean $E(Z_{i,t}) = \mu_i$ for each $i = 1, 2, \dots, m$. Then, the correlation matrix function for the vector process (Wei, 1989) is defined as following:

$$\Phi(k) = \text{Cor}\{\mathbf{Z}_t, \mathbf{Z}_{t+k}\} = \begin{pmatrix} \rho_{11}(k) & \rho_{12}(k) & \dots & \rho_{1m}(k) \\ \rho_{21}(k) & \rho_{22}(k) & \dots & \rho_{2m}(k) \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{m1}(k) & \rho_{m2}(k) & \dots & \rho_{mm}(k) \end{pmatrix} = \text{Cor}\{\mathbf{Z}_{t-k}, \mathbf{Z}_t\} \quad (7)$$

where $\rho_{ij}(k)$ is the cross-correlation between $Z_{i,t}$ and $Z_{j,t+k}$ for $k = 0, \pm 1, \pm 2, \dots$, $i, j = 1, 2, \dots, m$. For $i = j$, $\rho_{ii}(k)$ is the autocorrelation function for the i th component process $Z_{i,t}$ with lag k ; for $i \neq j$, $\rho_{ij}(k)$ is the cross-correlation function between the processes $Z_{i,t}$ and $Z_{j,t}$ with lag k . In this chapter, $\rho_{ij}(k)$ is expressed by the circular correlation coefficient as defined in (3). Note that, for $k \neq 0$, the correlation matrix function is not symmetric, since $\rho_{ij}(k) \neq \rho_{ji}(k)$ for $i \neq j$ in general.

Assuming that these m (stationary) component processes, $Z_{1,t}, Z_{2,t}, \dots, Z_{m,t}$, are connected to each other (clockwise) around a circle, we have a relationship between these random processes (see Figure 2). This idea can be explained by the following example. December 1999 is in the year 1999, whereas January 2000 belongs to the year

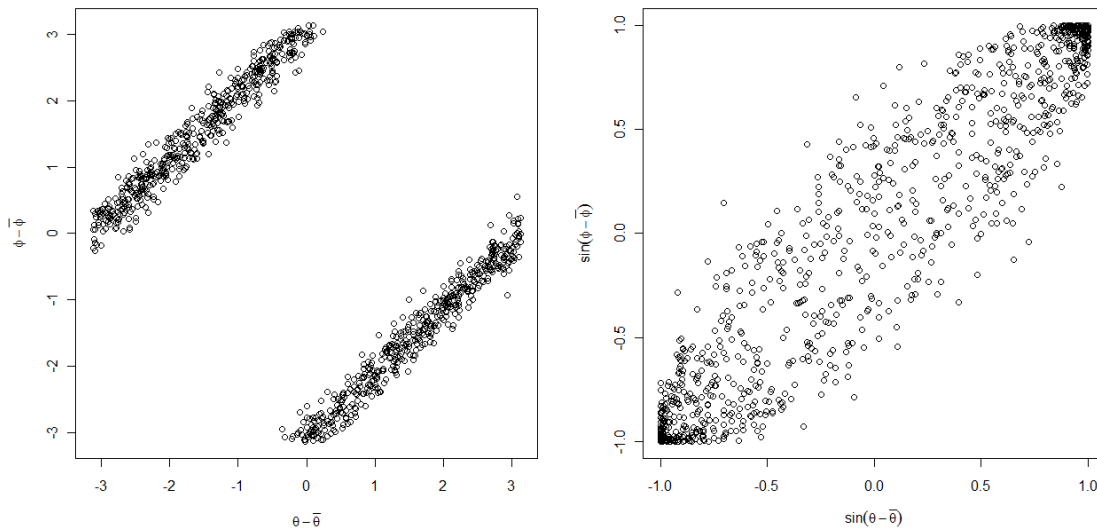


FIGURE 1. The data is simulated from the bivariate cosine distribution with low concentration but high positive correlation shown on the left. The sin transformation is applied to the data, and produce the scatter plot on the right

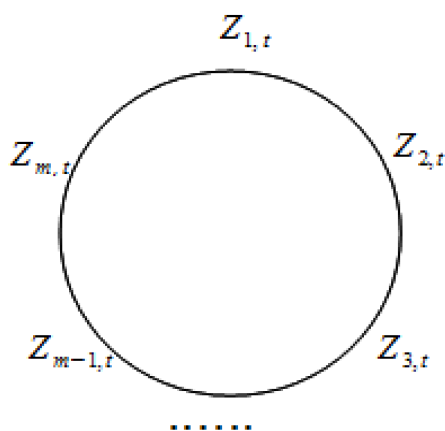


FIGURE 2. The component processes $Z_{1,t}, Z_{2,t}, \dots, Z_{m,t}$ are connected to each other clockwise around a circle.

2000. However, December 1999 is followed immediately by January 2000. In general, some repeated blocks of twelve different months are connected to each other around a circle, although these months belong to different year blocks. We then construct a correlation matrix function for a vector process having such a relationship on a circle.

If any two processes, $Z_{i,t}$ and $Z_{j,t}$, are apart from each other less than $m/2$, the cross-correlation $\rho_{ij}(0)$ between $Z_{i,t}$ and $Z_{j,t}$ is calculated. Otherwise, if $i < j$, we compute the cross-correlation $\rho_{ij}(-1)$ for the nearest two process of $Z_{i,t}$ and $Z_{j,t}$ with a lag -1 , (i.e, $Z_{j,t-1}$). If $i > j$, we compute the cross-correlation $\rho_{ij}(1)$ for the nearest two process of $Z_{i,t}$ and $Z_{j,t}$ with a lag 1 , (i.e., $Z_{j,t+1}$). This gives a special case of the circular correlation matrix function for the vector process:

$$\Psi = \text{Cor}\{\mathbf{Z}_t, \mathbf{Z}_{t+\mathbf{k}}\} = \begin{pmatrix} \rho_{11}(k) & \rho_{12}(k) & \dots & \rho_{1m}(k) \\ \rho_{21}(k) & \rho_{22}(k) & \dots & \rho_{2m}(k) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{m1}(k) & \rho_{m2}(k) & \dots & \rho_{mm}(k) \end{pmatrix} \quad (8)$$

where $k = 0$ for all $|i - j| \leq m/2$; Otherwise, $k = -1$ if $i < j$ while $k = 1$ if $i > j$; $i, j = 1, \dots, m$. Note that this correlation matrix is symmetric, and $\rho_{ab}(-1) = \rho_{ba}(1)$ for any pair of a, b .

The sample circular correlation matrix functions of $\Phi(k)$ and $\Psi(k)$ is obtained by replacing ρ_{ij} in (7) by r_{ij} for all $i, j = 1, 2, \dots, m$. This gives

$$\hat{\Phi}(k) = \hat{\text{COr}}\{\mathbf{Z}_t, \mathbf{Z}_{t+\mathbf{k}}\} = \begin{pmatrix} r_{11}(k) & r_{12}(k) & \dots & r_{1m}(k) \\ r_{21}(k) & r_{22}(k) & \dots & r_{2m}(k) \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1}(k) & r_{m2}(k) & \dots & r_{mm}(k) \end{pmatrix} = \hat{\text{COr}}\{\mathbf{Z}_{t-\mathbf{k}}, \mathbf{Z}_t\} \quad (9)$$

for $k = 0, \pm 1, \pm 2, \dots, i, j = 1, 2, \dots, m$, and

$$\hat{\Psi} = \hat{\text{COr}}\{\mathbf{Z}_t, \mathbf{Z}_{t+\mathbf{k}}\} = \begin{pmatrix} r_{11}(k) & r_{12}(k) & \dots & r_{1m}(k) \\ r_{21}(k) & r_{22}(k) & \dots & r_{2m}(k) \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1}(k) & r_{m2}(k) & \dots & r_{mm}(k) \end{pmatrix} \quad (10)$$

where $k = 0$ for all $|i - j| \leq m/2$; Otherwise, $k = -1$ if $i < j$ while $k = 1$ if $i > j$; $i, j = 1, \dots, m$.

4. AN EXAMPLE

The RNA data, used in Frellsen *et al.* (2009), consists of 267 pieces of RNA, each of which is different length and contains a repeated sequence of 7 dihedral angles. This data is represented by a time series of dihedral angles as follows:

$$\dots, \delta_{i-1}, \varepsilon_{i-1}, \zeta_{i-1}, \alpha_i, \beta_i, \gamma_i, \chi_i, \delta_i, \varepsilon_i, \zeta_i, \alpha_{i+1}, \beta_{i+1}, \gamma_{i+1}, \dots \quad (11)$$

where $i = 2, 3, \dots, n - 1$; $n = 8202$ indicates total number of nucleotides, each of which is a block of 7 dihedral angles. (11) can be re-written as a data matrix of n rows and 7

columns:

$$\mathbf{A}_t = \begin{pmatrix} \alpha_1 & \beta_1 & \gamma_1 & \chi_1 & \delta_1 & \varepsilon_1 & \zeta_1 \\ \alpha_2 & \beta_2 & \gamma_2 & \chi_2 & \delta_2 & \varepsilon_2 & \zeta_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha_n & \beta_n & \gamma_n & \chi_n & \delta_n & \varepsilon_n & \zeta_n \end{pmatrix}. \quad (12)$$

Figure 3 gives a length distribution of the pieces of RNA. The average length of the dihedral angles per piece is around 224, which corresponds to 32 nucleotides in length. Most of the pieces of RNA are not beyond 200 dihedral angles in length, and the longest piece has 2100 dihedral angles in length.

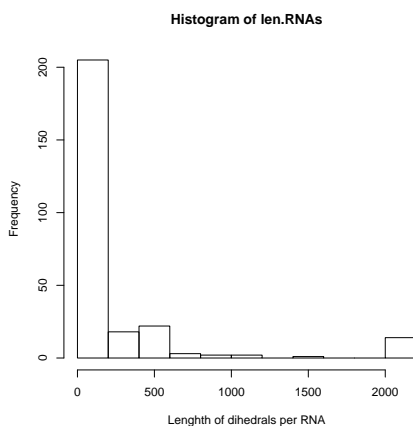


FIGURE 3. The length distribution of the pieces of RNA

Murray *et al.* (2003) found that the χ angle is less dependent on the backbone dihedral angles. Their preliminary work showed that the amplitude of sugar pucker seems quite constant, and for RNA the only two puckers that occur in high-quality data are $C3'$ endo and $C2'$ endo, they could use the δ dihedral angle alone to describe sugar pucker rather than the two variables of phase and amplitude. Consequently, they included the six backbone dihedral angles, α , β , γ , δ , ε and ζ , in their work.

In this example, we would like to investigate the correlation of the six backbone dihedral angles only. The data matrix A is then reduced to

$$\mathbf{A}_t = \begin{pmatrix} \alpha_1 & \beta_1 & \gamma_1 & \delta_1 & \varepsilon_1 & \zeta_1 \\ \alpha_2 & \beta_2 & \gamma_2 & \delta_2 & \varepsilon_2 & \zeta_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha_n & \beta_n & \gamma_n & \delta_n & \varepsilon_n & \zeta_n \end{pmatrix}, \quad (13)$$

Each row of \mathbf{A}_t represents a nucleotide division, whereas each column indicates the different types of dihedral angles. First of all, the sample circular correlation is computed for each pair of the dihedral angles using (3), shown on the upper diagonal entries of Figure 4. Some correlation coefficients are large in absolute value, e.g., $\text{Cor}(\alpha, \gamma) = -0.40$ and $\text{Cor}(\delta, \varepsilon) = 0.56$. And also, histogram of the each dihedral

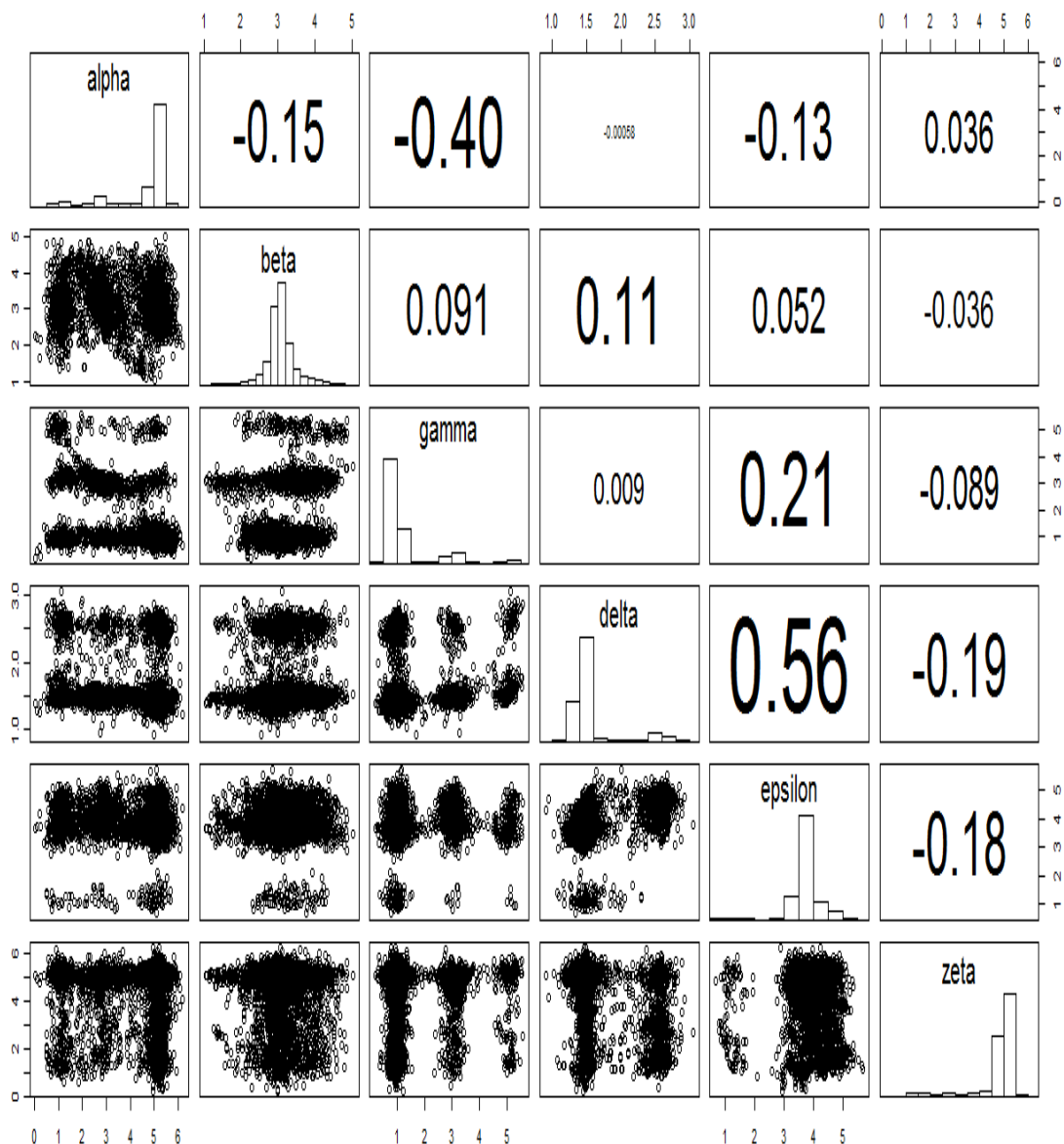


FIGURE 4. Pairwise circular correlation plots

angle is plotted on the diagonal of the figure. Secondly, for each dihedral angle, the autocorrelation function (ACF) is calculated as the lag increases. See the first 6 correlograms in Figure 5. The y-axis indicates the autocorrelation function, whereas the x-axis represents the lag, k . The significance test of the circular correlation coefficient is performed and the red points indicate which ACFs are significant at level 0.01. For

these two angles δ and ε , shown on the second row of the figure, their ACFs with lag 1 (and 2) have large values. This imply that each of the these angles is correlated to itself at lag 1 (and 2). Further, suppose the data is represented by a time series as in (11), then the ACF is calculated with different lags, shown in the last correlogram of the figure. It can be observed that the ACFs are periodic as the lag increases, and they are all significant. So there exists correlations across the time series.

Let $\mathbf{A}_t = [\alpha_t, \beta_t, \gamma_t, \delta_t, \varepsilon_t, \zeta_t]$ be an m-dimensional jointly stationary real-valued vector time series with sample mean directions $\bar{\mathbf{A}}_t = [\bar{\alpha}_t, \bar{\beta}_t, \bar{\gamma}_t, \bar{\delta}_t, \bar{\varepsilon}_t, \bar{\zeta}_t]$. The sample correlation matrix function, $\hat{\Phi}(k)$, is then calculated based on \mathbf{A}_t with the lag $k = 1, 2, \dots, 5$. In addition, we test if each circular correlation coefficient in the each $\hat{\Phi}(k)$ is significant by calculating its P-value of the underlying distribution (5). In general, we found that these P-values become less significant if the lag increases. $\hat{\Phi}(k)$ reports large values when the lag $k = 1, 2$. These $\hat{\Phi}(k)$ are given in Appendix A together with their corresponding matrices of the P-values.

Both the $\hat{\Phi}(0)$ and $\hat{\Psi}$ are calculated together with their corresponding P-values:

$$\hat{\Phi}(0) = \begin{pmatrix} & \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \\ \alpha & 1.000 & -0.151 & -0.397 & -0.001 & -0.133 & 0.036 \\ \beta & -0.151 & 1.000 & 0.091 & 0.110 & 0.052 & -0.036 \\ \gamma & -0.397 & 0.091 & 1.000 & 0.009 & 0.211 & -0.089 \\ \delta & -0.001 & 0.110 & 0.009 & 1.000 & 0.559 & -0.193 \\ \varepsilon & -0.133 & 0.052 & 0.211 & 0.559 & 1.000 & -0.179 \\ \zeta & 0.036 & -0.036 & -0.089 & -0.193 & -0.179 & 1.000 \end{pmatrix}$$

with its P-values

$$\begin{pmatrix} & \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \\ \alpha & 0 & 0 & 0 & 0.9697 & 0 & 0.0071 \\ \beta & 0 & 0 & 0 & 0 & 0.0004 & 0.0108 \\ \gamma & 0 & 0 & 0 & 0.5034 & 0 & 0 \\ \delta & 0.9697 & 0 & 0.5034 & 0 & 0 & 0 \\ \varepsilon & 0 & 0.0004 & 0 & 0 & 0 & 0 \\ \zeta & 0.0071 & 0.0108 & 0 & 0 & 0 & 0 \end{pmatrix};$$

$$\hat{\Psi} = \begin{pmatrix} & \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \\ \alpha & 1.000 & -0.151 & -0.397 & -0.001 & 0.151 & -0.154 \\ \beta & -0.151 & 1.000 & 0.091 & 0.110 & 0.052 & -0.023 \\ \gamma & -0.397 & 0.091 & 1.000 & 0.009 & 0.211 & -0.089 \\ \delta & -0.001 & 0.110 & 0.009 & 1.000 & 0.559 & -0.193 \\ \varepsilon & 0.151 & 0.052 & 0.211 & 0.559 & 1.000 & -0.179 \\ \zeta & -0.154 & -0.023 & -0.089 & -0.193 & -0.179 & 1.000 \end{pmatrix}$$

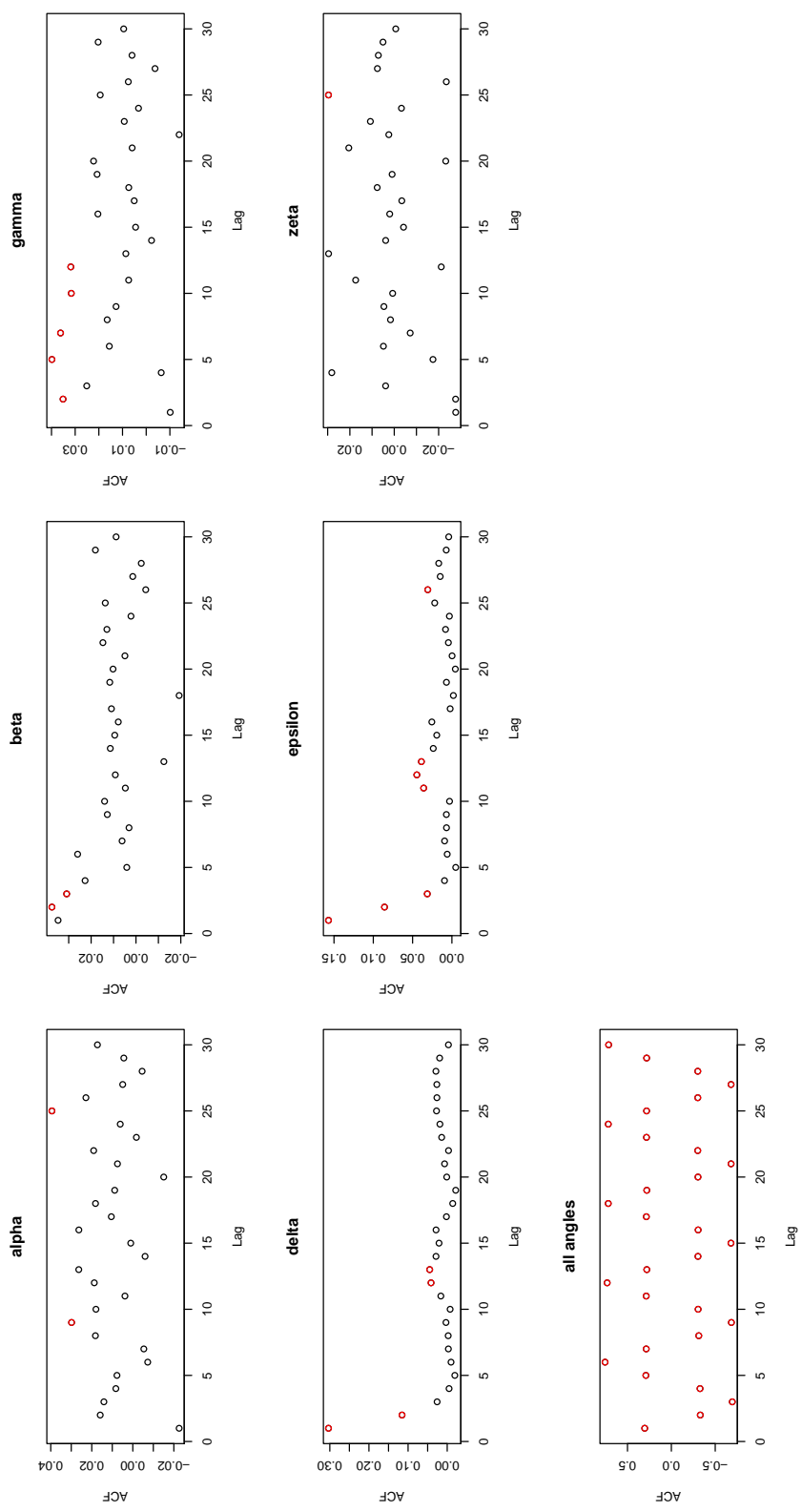


FIGURE 5. Correlograms for each of 6 dihedral angles and for all the angles

with its P-values

$$\begin{pmatrix} & \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \\ \alpha & 0 & 0 & 0 & 0.9697 & 0 & 0 \\ \beta & 0 & 0 & 0 & 0 & 0.0004 & 0.1822 \\ \gamma & 0 & 0 & 0 & 0.5034 & 0 & 0 \\ \delta & 0.9697 & 0 & 0.5034 & 0 & 0 & 0 \\ \varepsilon & 0 & 0.0004 & 0 & 0 & 0 & 0 \\ \zeta & 0 & 0.1822 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

These two correlation matrix functions have 6 entries different, both on the top right and the bottom left corners. Assuming independence of each pair of the correlation coefficients, we use Corollary 1 and two-sample z -test for testing equality of each of these pairs. For $\text{Cor}(\alpha, \varepsilon)$, under the null hypothesis $H_0 : \rho_{\alpha, \varepsilon}^{(\Psi)} = \rho_{\alpha, \varepsilon}^{(\Phi(0))}$, we have a test

statistic $|\sqrt{n-1}r_{\alpha, \varepsilon}^{(\Psi)} - \sqrt{n}r_{\alpha, \varepsilon}^{(\Phi(0))}| \sqrt{\frac{\hat{\lambda}_{20}^{(\Psi)} \hat{\lambda}_{02}^{(\Psi)}}{\hat{\lambda}_{22}^{(\Psi)}} + \frac{\hat{\lambda}_{20}^{(\Phi(0))} \hat{\lambda}_{02}^{(\Phi(0))}}{\hat{\lambda}_{22}^{(\Phi(0))}}} = 26.387$, which is tested

against the 5% critical value 1.96 provided by the standard normal distribution. The corresponding p -value is close to 0, so we are not going to accept the null hypothesis that $\rho_{\alpha, \varepsilon}^{(\Psi)} = \rho_{\alpha, \varepsilon}^{(\Phi(0))}$. Similarly, for $\text{Cor}(\alpha, \zeta)$, we have a test statistic value 18.369, which corresponds to a very small P-value around 0, so we are not going to accept the null hypothesis that $\rho_{\alpha, \zeta}^{(\Psi)} = \rho_{\alpha, \zeta}^{(\Phi(0))}$. For $\text{Cor}(\beta, \zeta)$, we are not going to reject the null hypothesis that $\rho_{\beta, \zeta}^{(\Psi)} = \rho_{\beta, \zeta}^{(\Phi(0))}$ since we have a test statistic value, 1.241, which corresponds to a very small P-value 0.2145.

We prefer $\hat{\Psi}$ since the matrix measures the relationship of these dihedral angles around a circle, that is, the time series of (11) wraps around the circle in Figure 2. For $\text{Cor}(\alpha, \zeta)$, ζ_{i-1} is followed by α_i , so it is preferable to calculate the sample correlation coefficient, $r_{\alpha, \zeta}$, with lag 1.

5. CIRCULAR PRINCIPAL COMPONENT ANALYSIS

Let \mathbf{x} be a random vector and \mathbf{w} be a vector of positive numbers with $\sum_i w_i^2 = 1$. Hotelling developed a technique called standard linear combination (SLC), cited by (Mardia *et al.*, 1979, p 213), which is a linear combination $\mathbf{w}'\mathbf{x}$. The first principal component seeks a SLC of random variables which has the largest variance, the second principal component with the second largest variance, and so on. The first few components often contain important information of the data, and then can summarize the data. Usually the last few components can be ignored if they are less informative, i.e., count for a small proportion of variation. The components with the large variances are of great interest.

5.1. The population case. Principal component analysis (PCA) performs ‘an orthogonal transformation which transforms any set of variables into a set of new variables which are uncorrelated with each other’ (Mardia *et al.*, 1979, p 214). In this section, we will discuss new techniques of applying the principal component analysis to angular data. Each of these techniques involves a transformation of any angular vector from a p fold torus space into a new metric coordinate space, where the distance between any two points is well defined.

Definition 1. If \mathbf{x} is a random vector with mean $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix Σ , then the principal component transformation is defined as:

$$\mathbf{x} \rightarrow \mathbf{y} = \Gamma'(\mathbf{x} - \boldsymbol{\mu}) = \Gamma'\mathbf{x}, \quad (14)$$

where Γ is orthogonal, and $\Gamma'\Sigma\Gamma = \Lambda$ is a diagonal matrix with the eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ on the diagonal. The eigenvalues λ_i are strictly positive if Σ is positive definite. Then, the j -th principal component of \mathbf{x} can be expressed as the j -th element of \mathbf{y} :

$$y_j = \gamma'_{(j)}\mathbf{x}, \quad (15)$$

where $\gamma'_{(j)}$ is the j -th column of Γ and also the j -th largest eigenvector of Σ . Sometimes $\gamma'_{(j)}$ may be called the j -th principal component loadings.

	Transformation
The Angular Circular PCA	$\mathbf{x}_j^{[1]} = [(\theta_j - \mu_j - \pi) \bmod 2\pi] - \pi$, for all $j = 1, \dots, p$.
The Sine Circular PCA	$\mathbf{x}_j^{[2]} = \sin(\theta_j - \mu_j)$, for all $j = 1, \dots, p$.
The Euclidean Circular PCA	$\mathbf{x}_j^{[3]} = (\sin(\theta_j), \cos(\theta_j))$, for all $j = 1, \dots, p$.
The Complex Circular PCA	$\mathbf{x}_j^{[4]} = \exp\{i(\theta_j - \mu_j)\}$, for all $j = 1, \dots, p$.

TABLE 1. Descriptions of the four CPCA methods

Let $\boldsymbol{\theta} = \{\theta_j : j = 1, 2, \dots, p\}$ be a random vector of a multivariate circular distribution, and let $\boldsymbol{\mu} = \{\mu_j : j = 1, 2, \dots, p\}$ be the mean direction of $\boldsymbol{\theta}$. The following

transformations, summarized in Table 1, are applied to this random vector. First of all, let

$$\mathbf{x}_j^{[1]} = [(\theta_j - \mu_j - \pi) \bmod 2\pi] - \pi, \quad (16)$$

for all $j = 1, \dots, p$. We called this transformation *angular method*. Clearly, the resulting random variable, $\mathbf{x}_j^{[1]}$, lies in $[-\pi, \pi]$. Secondly, the random vector θ is corrected by taking its *mean direction*, and then taking the sine transformation. We called it, *sin method*, and defined it mathematically as:

$$\mathbf{x}_j^{[2]} = \sin(\theta_j - \mu_j), \quad (17)$$

for all $j = 1, \dots, p$. Further, Mu *et al.* (2005) proposed a method, referred to as *Euclidean method*, that was used to study protein dihedral angles. Their method took both the sine and cosine transformations, but did not shift its *mean direction* to the zero direction inside trigonometric functions. However, we propose:

$$\mathbf{x}_j^{[3]} = (\sin(\theta_j - \mu_j), \cos(\theta_j - \mu_j)). \quad (18)$$

for all $j = 1, \dots, p$. Note that the Euclidean method is invariant if the mean direction is shifted to $\mathbf{0}$ or not, and the corresponding principal components are interchangeable up to a rotation (the proof is in Appendix B). Recently, Altis *et al.* (2007) introduced a complex version of the Euclidean method, referred to as *complex PCA*, in which p angular variables naturally lead to p eigenvalues and eigenvectors although they are all complex. Mathematically, the method transforms θ_j to a complex number using

$$\mathbf{x}_j^{[4]} = \exp\{i(\theta_j - \mu_j)\}, \quad (19)$$

for all $j = 1, \dots, p$.

After doing either (16), (17), (18) or (19), a new set of random variables is produced, namely as, $\mathbf{x}^{[1]}$, $\mathbf{x}^{[2]}$, $\mathbf{x}^{[3]}$ or $\mathbf{x}^{[4]}$. The new random vector, $\mathbf{x}^{[i]}$, is then substituted into (14). Thus, the eigenvectors (and eigenvalues) of the covariance matrix can be calculated in the same way as the standard PCA. Note that (18) and (17) transform the variables into a different metric coordinate space built up by the trigonometric function(s), and also $x^{[1]}$, $x^{[2]}$ and $x^{[4]}$ are of length p but $x^{[3]}$ is of length $2p$. Further, the covariance matrix corresponding to $x^{[4]}$ is defined, in Altis *et al.* (2007), as:

$$C_{mn} = \mathbb{E}((\mathbf{x}_m^{[4]} - \mathbb{E}(\mathbf{x}_m^{[4]}))(\overline{\mathbf{x}_n^{[4]} - \mathbb{E}(\mathbf{x}_n^{[4]})})), \quad (20)$$

with $m, n = 1, \dots, p$ and $\overline{x_j^{[4]}}$ being the complex conjugate of $x_j^{[4]}$. The covariance matrix, C , is a Hermitian matrix with p real-valued eigenvalues μ_n and p complex eigenvectors $\mathbf{w}^{(n)}$,

$$C\mathbf{w}^{(n)} = \mu_n\mathbf{w}^{(n)}, \quad (21)$$

where the eigenvectors are unique up to an arbitrary constant θ_0 . Then, the complex principal components are defined as:

$$W_n = \mathbf{w}^{(n)}\mathbf{x}^{[4]} = r_n \exp\{i(\theta_n + \theta_0)\}. \quad (22)$$

Thus, the complex principal components are represented by their weights r_n and angles θ_n in (22).

5.2. The sample case. Consider the sample-based principal component analysis. Let

$$X = (\mathbf{x}_1, \dots, \mathbf{x}_n)' \quad (23)$$

be a $n \times p$ data matrix, and let \mathbf{a} be a standardized vector of a length p . Then $X\mathbf{a}$ gives n observations, on a new variable, each of which is a weighted sum of the columns of X . The sample covariance of this new variable is $\mathbf{a}'S\mathbf{a}$, where S is the sample covariance matrix of X . Such SLC with the largest variance turns out to be the first principal component of X . The sample version of Definition 1 is given as follows:

Definition 2. Let $X = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$ be a $n \times p$ data matrix with the sample mean $\bar{\mathbf{x}}$ and the sample covariance matrix S . Without loss of generality, let $\bar{\mathbf{x}} = \mathbf{0}$. Principal component transformation is defined by direct analogy with (14) as

$$Y = (X - \mathbf{1}\bar{\mathbf{x}})G, \quad (24)$$

where G is orthogonal, and $G'SG = \Lambda$ is a diagonal matrix with the eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ on the diagonal. These eigenvalues λ_j are strictly positive if S is positive definite. Then, the j -th principal component transformation of X is given by

$$\mathbf{y}_j = (X - \mathbf{1}\bar{\mathbf{x}})\mathbf{g}_{(j)}, \quad (25)$$

where $\mathbf{g}_{(j)}$ is the j -th column of G , and also the j -th largest eigenvector of S . The r th element of \mathbf{y}_j , y_{rj} , represents the score of j th principal component on the r th observation. Further, Λ is the sample covariance matrix of Y , that is, the columns of Y are uncorrelated and the variance of $\mathbf{y}_{(j)}$ is λ_j .

Definition 3. Let $\theta = \{\theta_{rj} : r = 1, \dots, n, j = 1, \dots, p\}$ be a $n \times p$ circular data matrix measured in degree or radian. For the each column of θ , $\theta_{.j}$, its sample *mean direction* is defined in Mardia & Jupp (1999, p 15) as the following:

$$\bar{\theta}_j = \text{atan2}\left(\sum_r \sin \theta_{rj}, \sum_r \cos \theta_{rj}\right), \quad (26)$$

where r refers to the r -th observation of $\theta_{.j}$. Putting all mean directions, $\bar{\theta}_j$, together we have

$$\bar{\theta} = (\bar{\theta}_1, \dots, \bar{\theta}_p).$$

Assuming some circular data $\theta = \{\theta_{rj} : r = 1, \dots, n, j = 1, \dots, p\}$, then we would like to transform θ on the p fold torus onto X on the new coordinate metric space. For all $r = 1, \dots, n$ and $j = 1, \dots, p$, these $x_{rj}^{[1]}, x_{rj}^{[2]}, x_{rj}^{[3]}$ and $x_{rj}^{[4]}$ are calculated using Table 2 having a direct analogy with Table 1. After that, we apply the principal component analysis on any transformed data matrix X .

	Transformation
The Angular Circular PCA	$x_{rj}^{[1]} = [(\theta_{rj} - \bar{\theta}_j - \pi) \bmod 2\pi] - \pi.$
The Sine Circular PCA	$x_{rj}^{[2]} = \sin(\theta_{rj} - \bar{\theta}_j).$
The Euclidean Circular PCA	$x_{rj}^{[3]} = (\sin(\theta_{rj}), \cos(\theta_{rj})).$
The Complex Circular PCA	$x_{rj}^{[4]} = \exp\{i(\theta_{rj} - \bar{\theta}_j)\}.$

TABLE 2. Descriptions of the four CPCA methods

5.3. Display. The principal component analysis makes it possible to view the multi-dimensional data in the lower dimension, especially in 2-D. Usually we are looking for the first a few principal components that explain most of the variance. For example, if the first two components explain most of the variance, then a scattergram of the *scores* of these two components will often give a fair indication of the overall distribution of data (Mardia *et al.*, 1979, p 227). In summary, principal component analysis is a data reduction analysis in which the multidimensional data maps onto the first a few components which can explain the overall data.

6. CPCA FOR SINE MODEL

Let (θ, ϕ) be random variables with zero means. The density function, $f_s(\theta, \phi)$, is written as:

$$f_s(\theta, \phi) \propto \exp\{\kappa_1 \cos(\theta) + \kappa_2 \cos(\phi) - \lambda \sin(\theta) \sin(\phi)\}$$

for $-\pi \leq (\theta, \phi) < \pi$, where $\{\kappa_1, \kappa_2\} > 0$, $-\infty < \lambda < \infty$. In this section, if not specified, we assume that $\kappa_1 = \kappa_2$. So it is a particular example of exchangeable variables θ and ϕ .

6.1. The angular method and the sine method. In a plane, we have

$$\text{Cov}(x, y) = \text{Cov}(y, x)$$

as for the pdf $f(x, y) = f(y, x)$.

For the angular method, the two principal components are $\theta + \phi$ and $\theta - \phi$. If the data are concentrated (i.e., $\theta + \phi$ and $\theta - \phi$ are small), then $\sin(\theta + \phi) \approx \theta + \phi$ and $\sin(\theta - \phi) \approx \theta - \phi$. In this case, the sine method is approximately angular method.

For the sine method, consider $x = \sin(\theta + \phi)$, $y = \sin(\theta - \phi)$. To show $\mathbb{E}(xy) = 0$,

$$\mathbb{E}(\sin^2(\theta) \cos^2(\phi) - \cos^2(\theta) \sin^2(\phi)) = 0$$

which is true if θ, ϕ are exchangeable. Then, x, y work as the principal components.

For concentrated data, the sine method will not make a great difference from the angular method. In Section 7.1, the principal components in both the methods explain the data well.

6.2. The Euclidean method. Now consider the Euclidean case, let

$$x_1 = \cos(\theta), x_2 = \sin(\theta), x_3 = \cos(\phi), x_4 = \sin(\phi).$$

Then,

$$\Sigma = \begin{pmatrix} a & 0 & d & 0 \\ 0 & b & 0 & e \\ d & 0 & a & 0 \\ 0 & e & 0 & b \end{pmatrix}$$

where

$$\begin{aligned} a &= \text{Var}(\cos(\theta)) = \mathbb{E}(\cos^2(\theta)) - \mathbb{E}(\cos(\theta))^2; \\ b &= \text{Var}(\sin(\theta)) = \mathbb{E}(\sin^2(\theta)) - \mathbb{E}(\sin(\theta))^2 = 1 - \mathbb{E}(\cos^2(\theta)); \\ d &= \text{Cov}(\cos(\theta), \cos(\phi)) = \mathbb{E}(\cos(\theta)\cos(\phi)) - \mathbb{E}(\cos(\theta))\mathbb{E}(\cos(\phi)); \\ e &= \text{Cov}(\sin(\theta), \sin(\phi)) = \mathbb{E}(\sin(\theta)\sin(\phi)) - \mathbb{E}(\sin(\theta))\mathbb{E}(\sin(\phi)) = \mathbb{E}(\sin(\theta)\sin(\phi)); \\ 0 &= \text{Cov}(\cos(\theta), \sin(\theta)) = \mathbb{E}(\cos(\theta)\sin(\theta)) - \mathbb{E}(\cos(\theta))\mathbb{E}(\sin(\theta)) = \mathbb{E}(\sin(2\theta)); \\ 0 &= \text{Cov}(\cos(\phi), \sin(\phi)), \text{ analogously}; \\ 0 &= \text{Cov}(\cos(\theta), \sin(\phi)) = \text{Cov}(\sin(\theta), \cos(\phi)) \\ &= \mathbb{E}(\cos(\theta)\sin(\phi)) - \mathbb{E}(\cos(\theta))\mathbb{E}(\sin(\phi)) = \mathbb{E}(\cos(\theta)\sin(\phi)) [\text{See Appendix C}]. \end{aligned}$$

The the inverse of the covariance matrix is:

$$\Sigma^{-1} = \begin{pmatrix} \frac{a}{a^2-b^2} & 0 & \frac{-d}{a^2-b^2} & 0 \\ 0 & \frac{b}{b^2-e^2} & 0 & \frac{-e}{b^2-e^2} \\ \frac{-d}{a^2-d^2} & 0 & \frac{a}{a^2-d^2} & 0 \\ 0 & \frac{-e}{b^2-e^2} & 0 & \frac{b}{b^2-e^2} \end{pmatrix},$$

and its determinant can be written as:

$$\begin{aligned} |\Sigma^{-1}| &= b^2a^2 - e^2a^2 - b^2d^2 + e^2d^2 \\ &= b^2(a^2 - d^2) - e^2(a^2 - d^2) \\ &= (b^2 - e^2)(a^2 - d^2) \\ &= (b - e)(b + e)(a - d)(a + d). \end{aligned}$$

Thus, $(b - e)$, $(b + e)$, $(a - d)$ and $(a + d)$ are eigenvalues of Σ , which corresponds to

the eigenvectors, $\begin{pmatrix} 0 \\ -1 \\ 0 \\ 1 \end{pmatrix}$, $\begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$, $\begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}$, respectively.

For example, a random sample of 10000 data points are simulated from the sine distribution with $\kappa = (0.8, 0.8, 0.6)$ and $\mu = (0, 0)$, then the sample covariance matrix

can be calculated numerically:

$$S = \begin{pmatrix} 0.3932 & 0 & 0.0011 & 0 \\ 0 & 0.4775 & 0 & 0.1280 \\ 0.0011 & 0 & 0.3932 & 0 \\ 0 & 0.1280 & 0 & 0.4775 \end{pmatrix}.$$

A spectral decomposition of S yields principal components

$$\begin{aligned} y_1 &= 0.7071x_2 + 0.7071x_4, \\ y_2 &= 0.7071x_1 + 0.7071x_3, \\ y_3 &= 0.7071x_1 - 0.7071x_3, \\ y_4 &= -0.7071x_2 + 0.7071x_4, \end{aligned}$$

with eigenvalues 0.6055, 0.3943, 0.3921 and 0.3495, respectively. The first principal component has the largest variance which counts for 34.77 % of the total variance, and the last three count for 22.64%, 22.52% and 20.07%, respectively. Note that the last three have similar proportions. It can be seen from the first component that it represents sum of $\sin(\theta)$ and $\sin(\phi)$, while the second component is a sum of $\cos(\theta)$ and $\cos(\phi)$. The third component indicates difference between $\cos(\theta)$ and $\cos(\phi)$, whereas the fourth represents difference between $\sin(\phi)$ and $\sin(\theta)$.

Let us also consider some concentrated data. Generating a random sample of 10000 data points are simulated from the sine distribution with $\kappa = (10, 10, 9)$ and $\mu = (0, 0)$, then the sample covariance matrix can be calculated numerically:

$$S = \begin{pmatrix} 0.0269 & 0 & 0.0168 & 0 \\ 0 & 0.1853 & 0 & 0.1390 \\ 0.0168 & 0 & 0.0269 & 0 \\ 0 & 0.1390 & 0 & 0.1853 \end{pmatrix}.$$

A spectral decomposition of S yields principal components

$$\begin{aligned} y_1 &= 0.7071x_2 + 0.7071x_4, \\ y_2 &= -0.7071x_2 + 0.7071x_4, \\ y_3 &= 0.7071x_1 + 0.7071x_3, \\ y_4 &= 0.7071x_1 - 0.7071x_3, \end{aligned}$$

with eigenvalues 0.3243, 0.0463, 0.0437 and 0.0101, respectively. The first principal component has the largest variance which counts for 76.41 % of the total variance, and the last three count for 10.91%, 10.30% and 2.38%, respectively. Note that the first principal component is dominant in this case, and the last one is not informative. It can be seen from the first component that it represents sum of $\sin(\theta)$ and $\sin(\phi)$, while the second component is a difference between $\sin(\phi)$ and $\sin(\theta)$. The third component indicates sum of $\cos(\theta)$ and $\cos(\phi)$, whereas the fourth represents difference between $\cos(\theta)$ and $\cos(\phi)$.

If κ, μ of the bivariate sine distribution are given, the principal components can be calculated either from the population-based covariance matrix, Σ , or from the sample-based covariance matrix, S . For a large sample, the sample covariance matrix, S , is very close to the covariance matrix Σ . For concentrated data and the Euclidean method, the first principal component is dominant. If we remove the last two principal components, the method becomes the sine method in some sense. This is because the first two eigenvectors and eigenvalues are exactly the same as for the sine method. In Sections 7.1 and 7.2, the scattergrams for the sin method will be as the same as the scattergrams on the first two principal components for the Euclidean method. As the data becomes uniform, all the four principal components are equally weighted.

6.3. The complex PCA method.

6.3.1. *Eigen-decomposition.* Let (θ, ϕ) be random variables with zero means, and let $z_1 = \exp(i\theta)$ and $z_2 = \exp(i\phi)$. Then, the complex covariance matrix can be written as:

$$\Gamma = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix}$$

with

$$\begin{aligned} \gamma_{11} &= \text{Var}(z_1) = \mathbb{E}(z_1 z_1^*) - \mathbb{E}(z_1)\mathbb{E}(z_1^*) = 1 - (\mathbb{E}(\cos(\theta)))^2 \in \mathbb{R}; \\ \gamma_{22} &= \text{Var}(z_2) = \mathbb{E}(z_2 z_2^*) - \mathbb{E}(z_2)\mathbb{E}(z_2^*) = 1 - (\mathbb{E}(\cos(\phi)))^2 \in \mathbb{R}; \\ \gamma_{12} &= \gamma_{21} = \text{Cov}(z_1, z_2) = \mathbb{E}(z_1 z_2^*) - \mathbb{E}(z_1)\mathbb{E}(z_2^*) \\ &= \mathbb{E}(\cos(\theta - \phi)) - \mathbb{E}(\cos(\theta))\mathbb{E}(\cos(\phi)), \end{aligned}$$

where $*$ indicates the complex conjugate. Note that $\mathbb{E}(\sin(\theta)) = \mathbb{E}(\sin(\phi)) = 0$ and $\mathbb{E}(\sin(\theta)\cos(\phi)) = \mathbb{E}(\cos(\theta)\sin(\phi)) = 0$ (the proof is in Appendix C). Then, the eigenvalue decomposition of Γ will be $\Gamma = UDU^*$ with

$$U = \begin{pmatrix} \frac{\gamma_{11} - \gamma_{22} + \tau}{2\gamma_{12}} & -1 \\ 1 & \frac{\gamma_{11} - \gamma_{22} + \tau}{2\gamma_{12}} \end{pmatrix} \equiv \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}$$

where

$$\tau = \sqrt{\gamma_{11}^2 - 2\gamma_{11}\gamma_{22} + \gamma_{22}^2 + 4\gamma_{12}^2}.$$

The corresponding eigenvalues are

$$l_1 = (\gamma_{11} + \gamma_{22} + \tau)/2, \quad (27)$$

$$l_2 = (\gamma_{11} + \gamma_{22} - \tau)/2. \quad (28)$$

Hence, the principal components of complex PCA can be written as:

$$\text{PC}_1 = \alpha \exp(i\theta) + \beta \exp(i\phi), \quad (29)$$

$$\text{PC}_2 = -\beta \exp(i\theta) + \alpha \exp(i\phi). \quad (30)$$

Note that the complex principal components are invariant under rotation and multiplying by a real constant.

6.3.2. *Exchangeable variables.* Let θ and ϕ be exchangeable, i.e., $\kappa_1 = \kappa_2$. Then, the two principal components are

$$\begin{aligned} z &= \exp(i\theta) + \exp(i\phi), \\ w &= \exp(i\theta) - \exp(i\phi). \end{aligned}$$

Given the principal components, the principal angles are defined as:

$$\begin{aligned} \psi &= \text{Arg}(\exp\{i\theta\} + \exp\{i\phi\}) \\ &= \text{Arg}(\cos(\theta) + \cos(\phi) + i(\sin(\theta) + \sin(\phi))) \\ &= \text{atan2}(\sin(\theta) + \sin(\phi), \cos(\theta) + \cos(\phi)), \end{aligned}$$

and

$$\begin{aligned} \eta &= \text{Arg}(\exp\{i\phi\} - \exp\{i\theta\}) \\ &= \text{Arg}(\cos(\phi) - \cos(\theta) + i(\sin(\phi) - \sin(\theta))) \\ &= \text{atan2}(\sin(\phi) - \sin(\theta), \cos(\phi) - \cos(\theta)). \end{aligned}$$

Then,

$$\begin{aligned} \psi - \eta &= \text{Arg}(\bar{z}w) \\ &= \text{Arg}(2 \sin(\theta - \phi)i) \\ &= \begin{cases} \frac{\pi}{2}, & \text{if } \theta - \phi > 0; \\ \frac{3\pi}{2}, & \text{if } \theta - \phi < 0; \\ \text{undefined}, & \text{if } \theta = \phi. \end{cases} \end{aligned}$$

That is,

$$\eta = \psi + \frac{\pi}{2} \quad \text{or} \quad \eta = \psi + \frac{3\pi}{2}.$$

Hence, there are two lines on the plot of the angles (See Sections 7.1, 7.2 and 7.3).

On the other hand, the principal weights are defined as:

$$\begin{aligned} r_1 &= |\exp\{i\theta\} + \exp\{i\phi\}| \\ &= |1 + \exp\{i(\phi - \theta)\}| \\ &= \sqrt{2 + 2\cos(\lambda)}, \end{aligned}$$

and

$$\begin{aligned} r_2 &= |\exp\{i\phi\} - \exp\{i\theta\}| \\ &= |\exp\{i(\phi - \theta)\} - 1| \\ &= \sqrt{2 - 2\cos(\lambda)} \end{aligned}$$

where $\lambda = \phi - \theta$. Then,

$$r_1^2 + r_2^2 = 4, \quad r_1, r_2 > 0.$$

If we are plotting r_1 against r_2 , there is an arc, that is, the positive part of a circle.

6.3.3. *Concentrated data.* Let us consider some data from the sine model. In this case, the distribution can be approximated by the following bivariate normal

$$\begin{pmatrix} \theta \\ \phi \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right].$$

For concentrated data, the complex variables become

$$\begin{aligned} z_1 &\approx 1 + i\theta, \\ z_2 &\approx 1 + i\phi. \end{aligned}$$

Then, the covariances are

$$\begin{aligned} \gamma_{11} &= \text{Var}(z_1) = \mathbb{E}((1 + i\theta)(1 - i\theta)) - \mathbb{E}(1 + i\theta)\mathbb{E}(1 - i\theta) = \mathbb{E}(\theta^2) = 1; \\ \gamma_{22} &= \text{Var}(z_2) = \mathbb{E}((1 + i\phi)(1 - i\phi)) - \mathbb{E}(1 + i\phi)\mathbb{E}(1 - i\phi) = \mathbb{E}(\phi^2) = 1; \\ \gamma_{12} &= \text{Cov}(z_1, z_2) = \mathbb{E}((1 + i\theta)(1 - i\phi)) - \mathbb{E}(1 + i\theta)\mathbb{E}(1 - i\phi) \\ &= \mathbb{E}(\theta\phi); \\ \gamma_{21} &= \text{Cov}(z_2, z_1) = \mathbb{E}((1 + i\phi)(1 - i\theta)) - \mathbb{E}(1 + i\phi)\mathbb{E}(1 - i\theta) \\ &= \mathbb{E}(\theta\phi). \end{aligned}$$

Thus, the covariance matrix can be written as

$$\Gamma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

The characteristic equation is

$$\det(\Gamma - \lambda I) = (1 - \lambda)^2 - \rho^2 = 0$$

with solutions $\lambda = 1 \pm \rho$. The corresponding eigenvectors are

$$U = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

A random sample is simulated from the sine distribution with $\kappa = (10, 10, 9)$ and $\mu = (0, 0)$. Then, the covariance matrix can be calculated numerically:

$$C = \begin{pmatrix} .2223 & .1663 \\ .1663 & .2223 \end{pmatrix}$$

A spectral decomposition of S yields principal components

$$\begin{aligned} y_1 &= 0.7071x_1 + 0.7071x_2, \\ y_2 &= -0.7071x_1 + 0.7071x_2 \end{aligned}$$

with eigenvalues 0.3886 and 0.0560, respectively. The first principal component has the largest variance which counts for 87.40 % of the total variance, and the second has the rest. It can be seen from the first component that it represents the sum of x_1 and x_2 , while the second component is a representation of difference of x_2 and x_1 .

6.3.4. *Visualization.* Let us consider cases under the bivariate sine models. The cartesian representation $z = x + iy$ can be converted into its polar coordinates, i.e., $\arg(z)$ and $|z|$. Altis *et al.* (2007) suggested to visualize the complex principal components into both the angles and modulus. Explicitly, the argument of PC_1 plotted against the argument of PC_2 , whereas $|PC_1|$ is plotted against $|PC_2|$. Then, we showed that $|PC_1|^2 + |PC_2|^2$ is some constant. This means that a quarter of circle appeared on the first quadrant of the modulus plot. It may not contain any useful information about PCs. On the angle plot, the original angular data was transformed by

$$\arg(PC_1) = \text{atan2}(\alpha \cos(\theta) + \beta \cos(\phi), \alpha \sin(\theta) + \beta \sin(\phi)), \quad (31)$$

$$\arg(PC_2) = \text{atan2}(\alpha \cos(\phi) - \beta \cos(\theta), \alpha \sin(\phi) - \beta \sin(\theta)). \quad (32)$$

For complex PCs, we can always perform some arbitrary rotation in order to make the transformed data centred at the origin. Apart from some limited cases giving us two lines, the transformed data was reasonable well to recover the directions which have maximum variances.

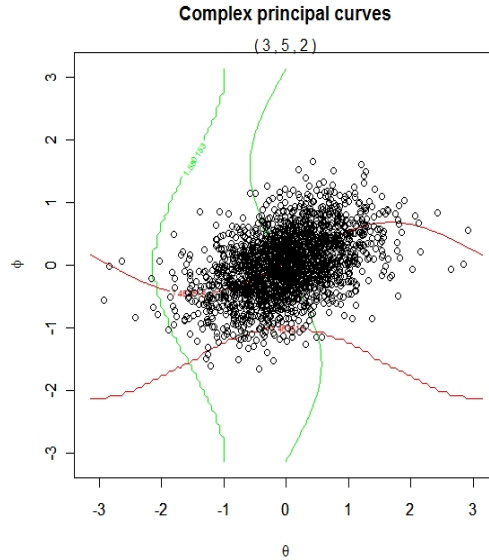


FIGURE 6. The principal curves

We investigated this transformation further by plotting some “principal curves” passing through mode of the data. Generating ten thousand data points from the sine distribution with $\kappa = (3, 5, 2)$ and $\mu = (0, 0)$, we then applied complex PCA to this data. As a result, the arguments of the two principal components are given by:

$$\arg(PC_1) = \text{atan2}(0.88 \cos(\theta) + 0.48 \cos(\phi), 0.88 \sin(\theta) + 0.48 \sin(\phi)), \quad (33)$$

$$\arg(PC_2) = \text{atan2}(0.88 \cos(\phi) - 0.48 \cos(\theta), 0.88 \sin(\phi) - 0.48 \sin(\theta)). \quad (34)$$

The corresponding eigenvalues are 0.4384 and 0.1651, respectively. It can be seen from Figure 6 that there are 4 level curves. Two of them (green) correspond to (33), whereas the others (red) use (34). The green curves have level value, 1.58, while the red curves are with level value, 1.46. In particular, the two curves passing through mode of the distribution are of interest. The curves (with different colors) are perpendicular to each other.

7. SIMULATION STUDIES

Ten thousand data points are simulated from the sine model with various κ parameters and zero mean. Ten thousand data points would be large enough to approximate their populations. In Sections 7.1, 7.2, 7.3 and 7.4, the κ parameters starts from (10, 10, 9), which corresponds to the concentrated data. The data becomes diffuse as the κ parameters decrease. At the end, the data is almost uniform. All circular PCA methods were applied to these data. For each method, eigenvectors and eigenvalues of sample covariance matrix are reported, and also scattergram of the first two principal components is drawn. In particular, for the complex PCA method, the principal angles and principal weights are drawn in the separate plots.

7.1. Concentrated data drawn from the sine distribution. Given $\kappa = (10, 10, 9)$ and $\mu = (0, 0)$, a random sample $\theta_i, \phi_i : r = 1, \dots, n$ is simulated from the bivariate sine distribution. The four different transformations are performed on the data set producing four different transformed data matrices, $X^{[1]}$, $X^{[2]}$, $X^{[3]}$ and $X^{[4]}$. Then, the principal component analysis is applied to each of these matrices.

For the angular method, the resulting principal components are

$$\begin{aligned} y_1 &= 0.7071x_1 + 0.7071x_2, \\ y_2 &= -0.7071x_1 + 0.7071x_2, \end{aligned}$$

with the corresponding variances, 0.3918 and 0.0581, respectively.

For the sine method, the resulting principal components are

$$\begin{aligned} y_1 &= 0.7071x_1 + 0.7071x_2, \\ y_2 &= -0.7071x_1 + 0.7071x_2, \end{aligned}$$

with the corresponding variances, 0.3251 and 0.0459, respectively.

For the Euclidean method, the resulting principal components are

$$\begin{aligned} y_1 &= 0.7071x_2 + 0.7071x_4, \\ y_2 &= 0.7071x_2 - 0.7071x_4, \\ y_3 &= -0.7071x_1 - 0.7071x_3, \\ y_4 &= 0.7071x_1 - 0.7071x_3, \end{aligned}$$

with the corresponding variances, 0.3251, 0.0459, 0.0240 and 0.0104, respectively.

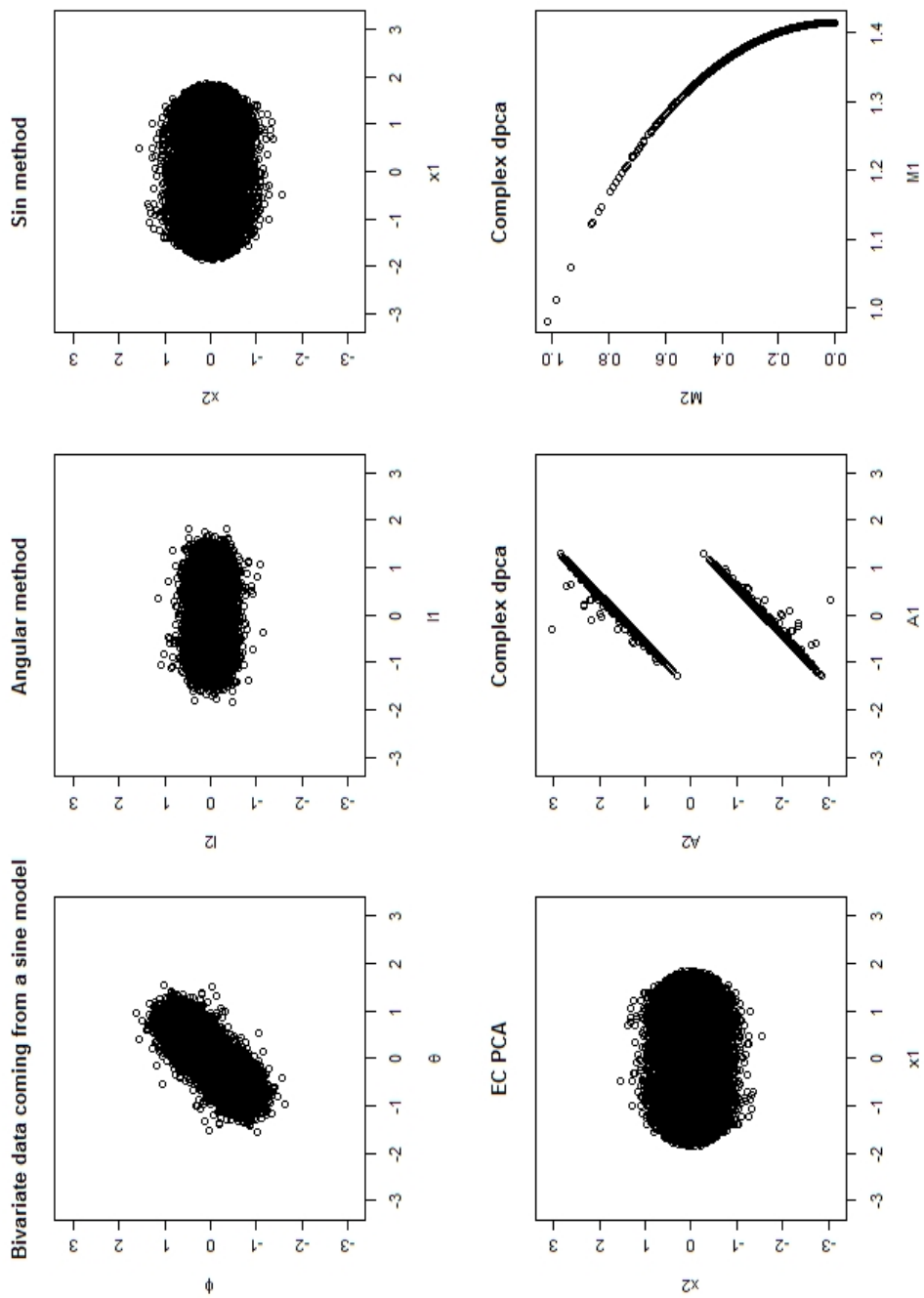


FIGURE 7

For the complex method, the resulting principal components are

$$y_1 = 0.7071x_1 + 0.7071x_2,$$

$$y_2 = 0.7071x_1 - 0.7071x_2$$

with the corresponding variances, 0.3886 and 0.0560, respectively.

Since $\kappa_1 = \kappa_2$, the random variables θ, ϕ are exchangeable. For some concentrated data, it can be approximated by some bivariate normal. All the circular PCA methods work well in this case. The first PC is dominated and the second one contributes only a small fraction of the total variation for all methods. For the angular method, the two PCs are $\theta + \phi$ and $\phi - \theta$, respectively. For the Euclidean method, the last two eigenvalues can be ignored since they count less information. The first two eigenvectors are $\sin(\theta) + \sin(\phi)$ and $\sin(\phi) - \sin(\theta)$ as same as ones obtained from the sine method. And also, the corresponding eigenvalues are roughly as same as ones obtained from the sine method. This means that the Euclidean method becomes the sine method as data is concentrated. For the complex method, the eigenvectors are all real. Then, the two PCs are $\exp(i\theta) + \exp(i\phi)$ and $\exp(i\theta) - \exp(i\phi)$, respectively.

The angular method performs translation and rotation of the original data, in which the resulting scores represent the maximal variations on two PCs. The sine and Euclidean methods gives an almost identical picture, which also can explain the directions of maximal variations.

7.2. Moderate data drawn from the sine distribution. Given $\kappa = (3, 3, 1)$ and $\mu = (0, 0)$, a random sample $\theta_i, \phi_i : r = 1, \dots, n$ is simulated from the bivariate sine distribution. The four different transformations are performed on the data set producing four different transformed data matrices, $X^{[1]}, X^{[2]}, X^{[3]}$ and $X^{[4]}$. Then, the principal component analysis is applied to each of these matrices.

For the angular method, the resulting principal components are

$$\begin{aligned} y_1 &= 0.7116x_1 + 0.7026x_2, \\ y_2 &= -0.7026x_1 + 0.7116x_2, \end{aligned}$$

with the corresponding variances, 0.5403 and 0.3384, respectively.

For the sine method, the resulting principal components are

$$\begin{aligned} y_1 &= 0.7121x_1 + 0.7021x_2, \\ y_2 &= -0.7021x_1 + 0.7121x_2, \end{aligned}$$

with the corresponding variances, 0.3472 and 0.2086, respectively.

For the Euclidean method, the resulting principal components are

$$\begin{aligned} y_1 &= 0.7121x_2 + 0.7021x_4, \\ y_2 &= 0.7021x_2 - 0.7121x_4, \\ y_3 &= -0.7437x_1 - 0.6686x_3, \\ y_4 &= -0.6686x_1 + 0.7437x_3, \end{aligned}$$

with the corresponding variances, 0.3472, 0.2086, 0.0727 and 0.0689, respectively.

For the complex method, the resulting principal components are

$$\begin{aligned} y_1 &= 0.7129x_1 + 0.7071x_2, \\ y_2 &= 0.7071x_1 - 0.7129x_2 \end{aligned}$$

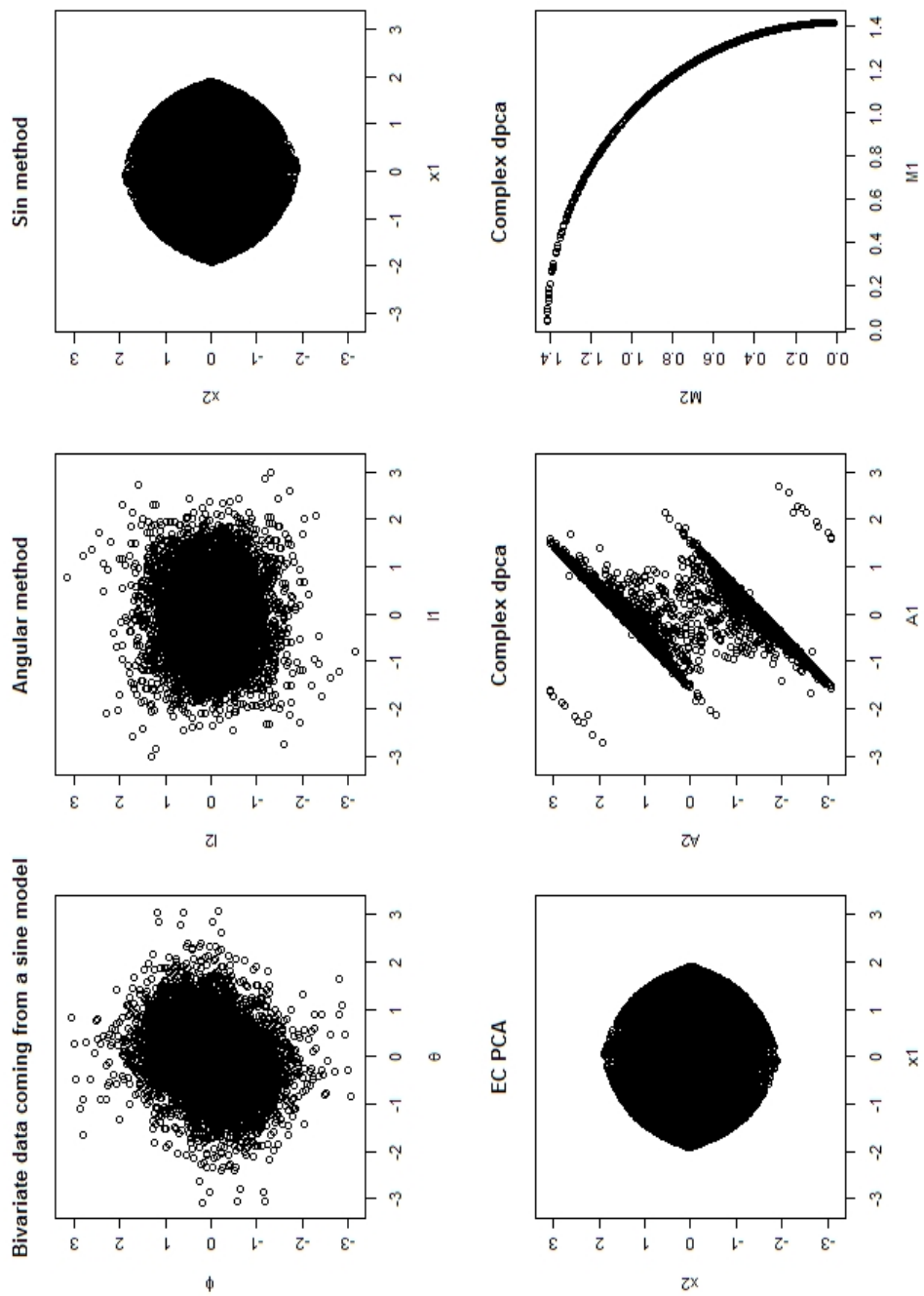


FIGURE 8

with the corresponding variances, 0.4199 and 0.2776, respectively.

For some moderate data, the second PC becomes important in comparison with the PC in Section 7.1. In other words, the first two eigenvectors count most of variation

than others. The PCs remain roughly the same as for the concentrated data in all the methods.

The angular method looks nice as for concentrated data. For the sine and Euclidean methods the resulting scores have a diamond shape on the first two PCs. This is because the sine transformation is not bijective. It will transform the two distinct angles to a point (e.g., $\sin(\pi/4) = \sin(-5\pi/4) = \sqrt{2}/2$). Thus, the pictures are not helpful for finding the directions of maximal variations.

7.3. Diffuse data drawn from the sine distribution. Given $\kappa = (0.8, 0.8, 0.6)$ and $\mu = (0, 0)$, a random sample $\theta_i, \phi_i : r = 1, \dots, n$ is simulated from the bivariate sine distribution. The four different transformations are performed on the data set producing four different transformed data matrices, $X^{[1]}$, $X^{[2]}$, $X^{[3]}$ and $X^{[4]}$. Then, the principal component analysis is applied to each of these matrices.

For the angular method, the resulting principal components are

$$\begin{aligned} y_1 &= 0.7094x_1 + 0.7049x_2, \\ y_2 &= -0.7049x_1 + 0.7094x_2, \end{aligned}$$

with the corresponding variances, 2.3445 and 1.4690, respectively.

For the sine method, the resulting principal components are

$$\begin{aligned} y_1 &= 0.7065x_1 + 0.7077x_2, \\ y_2 &= -0.7077x_1 + 0.7065x_2, \end{aligned}$$

with the corresponding variances, 0.6097 and 0.3344, respectively.

For the Euclidean method, the resulting principal components are

$$\begin{aligned} y_1 &= 0.7065x_2 + 0.7077x_4, \\ y_2 &= -0.7200x_1 - 0.6940x_3, \\ y_3 &= -0.6940x_1 + 0.7200x_3, \\ y_4 &= 0.7077x_2 - 0.7065x_4, \end{aligned}$$

with the corresponding variances, 0.6097 0.4105 0.3861 and 0.3344, respectively.

For the complex method, the resulting principal components are

$$\begin{aligned} y_1 &= 0.7076x_1 + 0.7066x_2, \\ y_2 &= 0.7066x_1 - 0.7076x_2 \end{aligned}$$

with the corresponding variances, 1.0202 and 0.7205, respectively.

For some diffuse data, the second PC becomes increasingly important. For the Euclidean method, the last two PCs cannot be excluded from the analysis since they contribute a fraction of total variation. And, the order of eigenvectors are different from the previous two examples. This method does not help with dimension reduction in this example, since the method transforms the data from the 2-D angular space into a 4-D Euclidean space, and then gives four PCs.

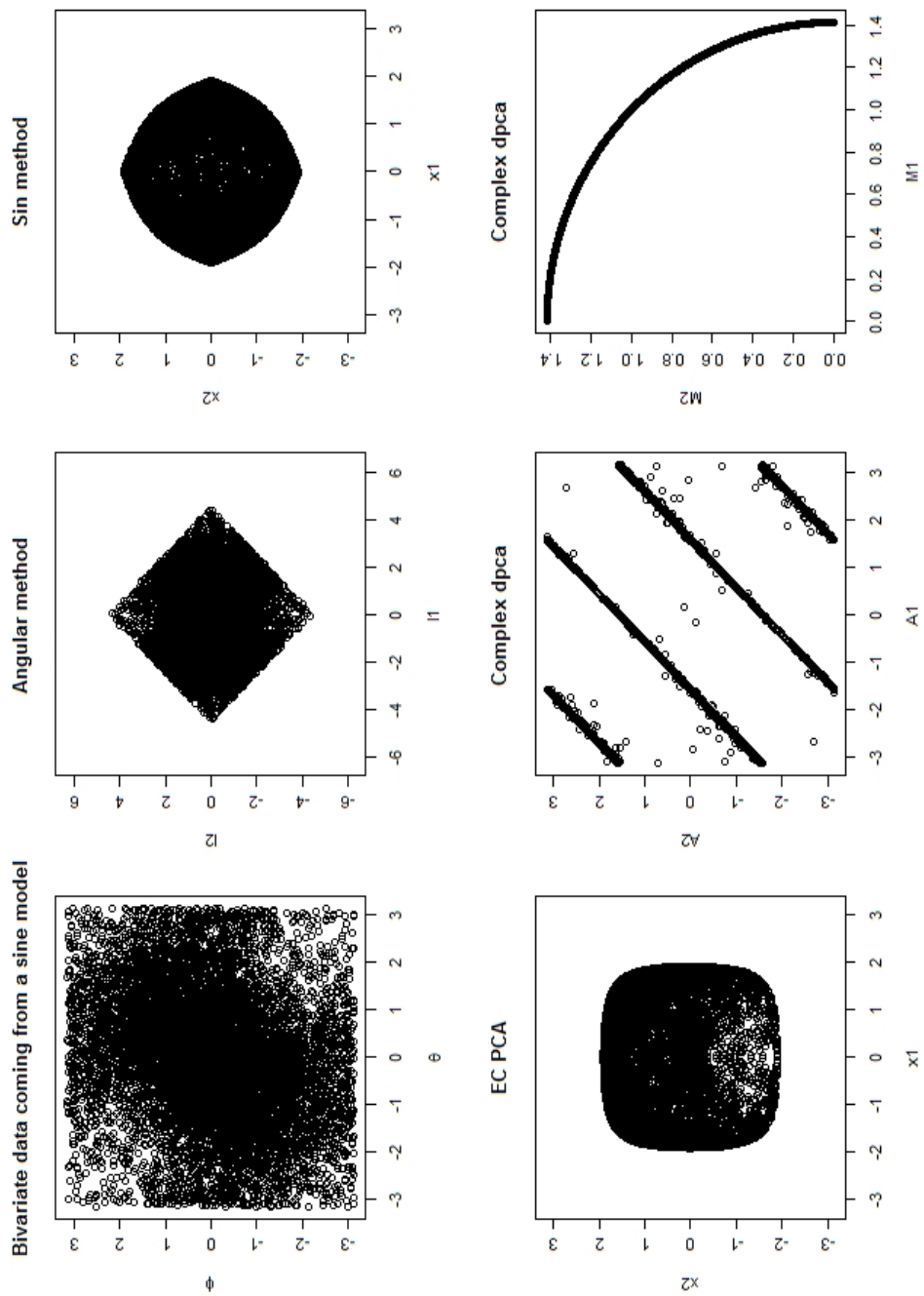


FIGURE 9

The angular method rotates the data by $\pi/4$. It makes no sense for these type of angular data, and the method should consider circularity. The sine and Euclidean methods are not working either. The Euclidean method gives us a full filled circle.

7.4. Semi-diffused data from the cosine distribution. Given $\kappa = (12, 0.5, 0.4)$ and $\mu = (0, 0)$, a random sample $\theta_i, \phi_i : r = 1, \dots, n$ is simulated from the bivariate cosine distribution. The four different transformations are performed on the data set producing four different transformed data matrices, $X^{[1]}$, $X^{[2]}$, $X^{[3]}$ and $X^{[4]}$. Then, the principal component analysis is applied to each of these matrices.

For the angular method, the resulting principal components are

$$\begin{aligned} y_1 &= -0.0149x_1 + 0.9999x_2, \\ y_2 &= 0.9999x_1 + 0.0149x_2, \end{aligned}$$

with the corresponding variances, 2.9724 and 0.0868, respectively.

For the sine method, the resulting principal components are

$$\begin{aligned} y_1 &= -0.0396x_1 + 0.9992x_2, \\ y_2 &= 0.9992x_1 + 0.0396x_2, \end{aligned}$$

with the corresponding variances, 0.4975 and 0.0800, respectively.

For the Euclidean method, the resulting principal components are

$$\begin{aligned} y_1 &= -0.0396x_2 + 0.9992x_4, \\ y_2 &= 0.0004x_1 + x_3, \\ y_3 &= 0.9992x_2 + 0.0396x_4, \\ y_4 &= x_1 - 0.0004x_3, \end{aligned}$$

with the corresponding variances, 0.4975, 0.4968, 0.0800 and 0.0032, respectively.

For the complex method, the resulting principal components are

$$\begin{aligned} y_1 &= 0.0179x_1 - 0.9998x_2, \\ y_2 &= 0.9998x_1 + 0.0179x_2 \end{aligned}$$

with the corresponding variances, 0.9939 and 0.0836, respectively.

Now we consider some semi-concentrated data from the cosine model, i.e., it is concentrated for one variable, but not another. Note that $\kappa_1 \neq \kappa_2$. This example gives us almost 1-D data in some sense. The first PC is dominated and the second one only counts a small fraction of the total variation for all methods except the Euclidean method. For angular method, the two PCs are $\alpha\phi - \beta\theta$ and $\alpha\theta + \beta\phi$, where $\alpha^2 + \beta^2 = 1$ and $\alpha \gg \beta$. They are approximately equal to ϕ and θ since β is small. Roughly speaking, the method only rotates the data by $\pi/2$ clockwise. For the sine method, the two PCs are $\alpha\sin(\phi) - \beta\sin(\theta)$ and $\alpha\sin(\theta) + \beta\sin(\phi)$, where $\alpha^2 + \beta^2 = 1$ and $\alpha \gg \beta$. They are approximately equal to $\sin(\phi)$ and $\sin(\theta)$ since β is small. For the Euclidean method, the first two PCs are almost equally weighted. The first two eigenvectors are $\alpha\sin(\phi) - \beta\sin(\theta)$ and $\alpha\cos(\phi) + \beta\cos(\theta)$, where $\alpha^2 + \beta^2 = 1$ and $\alpha \gg \beta$. They are approximately equal to $\sin(\phi)$ and $\cos(\phi)$ since β is small. For the complex method, the eigenvectors are all real. And, the two PCs are approximately $\exp(i\phi)$ and $\exp(i\theta)$, respectively.

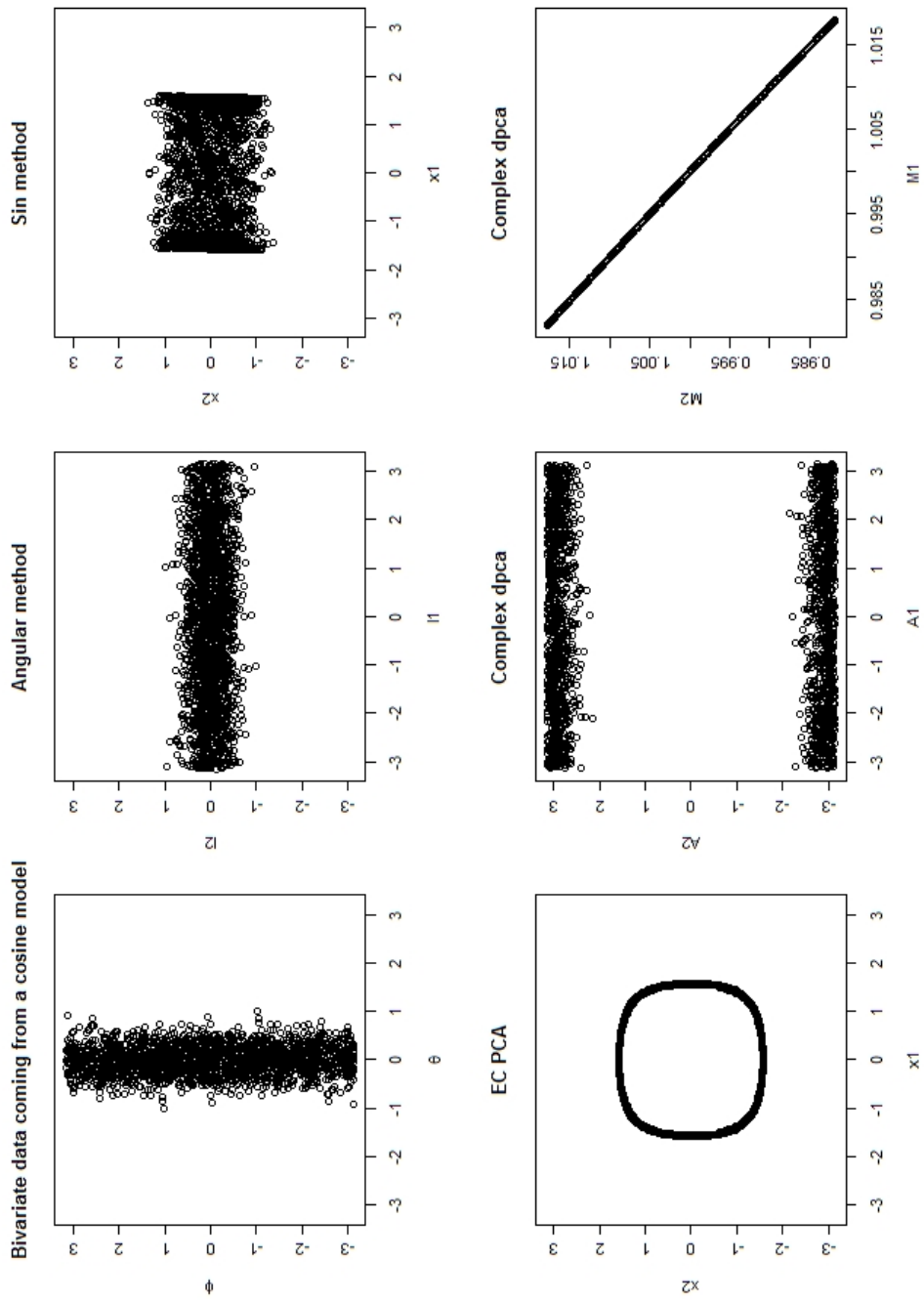


FIGURE 10

The cosine distribution would allow us to simulate almost 1-D data, i.e., very concentrated on a variable but diffused on the another. The angular method rotates the data by around $\pi/2$. Then, we can observe the direction of which the data varies the most. However, it may have problem at the edges ($-\pi$ and π). For instance, the upper middle panel of Figure 10 gives a scatter plot, in which the scores will not be continuous across the edges. For the sine method we plotted the scores $\sin(\phi)$ against $\sin(\theta)$ on the principal components. For Euclidean method, ϕ are wrapping around a circle by plotting $\cos(\phi)$ against $\sin(\phi)$.

Note that the complex method produces some lines on the principal angles for all data. They will not be useful. The principal weights just allow us to draw a quarter of circle. At the moment, we do not see usefulness of these pictures.

8. STRENGTHS AND WEAKNESSES

Let (θ, ϕ) be random variables with zero means. In this section, we assume some data $\theta_i, \phi_i, i = 1, \dots, n$ coming from the bivariate sine distribution with single mode (i.e., $\kappa_1 \kappa_2 > \lambda^2$).

8.1. The Angular Circular PCA. Given $\theta_i, \phi_i, i = 1, \dots, n$, we firstly standardize them by correcting their circular mean, and then make all data points lie between $-\pi$ and π . This method is simple and straightforward. It will pick out the directions which have maximum variances as standard PCA does. However, it is illegitimate to treat angles as numbers. It matters when data are diffuse. For example, $(179^\circ, 0)$ and $(-179^\circ, 0)$ would be two nearby points on the torus, but will be far away from each other if we consider angles as numbers. This is also true in general without assuming the sine distribution, and also this is a major drawback of the angular circular PCA.

8.2. The Sine Circular PCA. The sine circular PCA avoids the edge problems in the angular circular PCA by taking sine transformation of the standardized data. This method will map angles to some number between -2 and 2. Clearly the resulting principal scores do not involve angles any more. However, the principal scores will get folding as the angles beyond $\pi/2$ (and $-\pi/2$) due to nature of sine function. It makes these scores very hard to visualize in the scatter plot.

8.3. The Euclidean Circular PCA. The Euclidean circular PCA will take both sine and cosine transformations of the standardized data of two variables, and then gives four variables with sine and cosine. These four principal components can be written in terms of some linear combinations of sine and cosine. These principal components are also ordered by magnitudes of the corresponding eigenvalues. For different types of data (i.e., diffuse, moderate or concentrated), orders of the principal components change in accordance with magnitudes of the eigenvalues. Consider some concentrated data from the sin distribution with $\kappa = (10, 10, 9)$, for example. The first two principal component are more important than the other two. The first PC is dominated by the sum of sine terms, whereas the second PC is almost the difference of sine terms. Thus, the Euclidean circular PCA reduces to the sine circular PCA for concentrated data.

Note that the principal scores are not angles, and also they lie between -2 and 2. Also note that the number of principal components is doubled in comparison to the numbers in the other methods.

8.4. The Complex Circular PCA. The complex circular PCA represents angles in the complex plane using $\exp\{i\psi\}$, where $\psi = \theta$ or ϕ is standardized angular variable. Note that the complex covariance matrix is real under the above setting. However, it does not hold in general. Consider the following pair of variables with a phase change, $(\theta, \phi^* = \phi + \pi/2)$. In this case, $\mathbb{E}(\sin(\theta - \phi^*))$ would not be zero, then the off-diagonal entries of the covariance matrix are complex. Also note that the covariance matrices of both Euclidean circular PCA and complex circular PCA are built up by the circular moments of the sine distribution. Further, the complex principal components involve real parts and imaginary parts. The real part of the principal component consists of the linear combination of cosine terms, whereas the imaginary part of the principal component are in forms of sine terms. This means that this method treats the sine terms as equally important as the cosine terms. It is not true for the Euclidean circular PCA. For some limiting cases, the argument of the first principal component is parallel to the argument of the second principal component. The scatter plot of these arguments gives two lines which would not be meaningful for PCA analysis.

9. ANGULAR DISTANCE

Cluster analysis or clustering is known as an unsupervised pattern recognition (Theodoridis & Koutroumbas, 2009). Suppose that there exists a random sample X having a dimension $n \times p$. Each row represents a data point \mathbf{x}_i defined on the p -dimensional sample space. And then, the clustering algorithm is implemented to reveal the underlying similarities of X and cluster “similar” data points into groups.

Theodoridis & Koutroumbas (2009) discussed two major issues in cluster analysis. A major issue in cluster analysis is that of defining the “similarity” between two data points on a specific sample space and determining an appropriate distance measure for it. Another issue of importance is to choose an algorithm that will cluster the data into several groups on the basis of the adopted similarity measure. In general, the different clustering algorithms may lead to different results, and only the expert can tell which algorithm is more appropriate to be used on his or her data set.

Let $X = (\theta_{ij})$ be a data matrix, which defined on $[-\pi, \pi]$. Whereas $i \in 1, \dots, n$ represents i -th observation, $j \in 1, \dots, p$ indicates the j -th variable. For any pair of observations, $(\theta_{a.}, \theta_{b.})$, $a, b \in 1, \dots, n$, one choice of *angular distance* is defined:

$$d_1(\theta_{a.}, \theta_{b.}) = \sqrt{\sum_{j=1}^p \min(|\theta_{aj} - \theta_{bj}|, 2\pi - |\theta_{aj} - \theta_{bj}|)^2}. \quad (35)$$

Note that $d_1(\theta_{a.}, \theta_{b.})$ is not Euclidean. Alternatively, another choice of *angular distance* is defined as the Euclidean distance on the p -D torus:

$$\begin{aligned} d_2(\theta_{a.}, \theta_{b.}) &= \sqrt{\sum_{j=1}^p (\cos(\theta_{aj}) - \cos(\theta_{bj}))^2 + (\sin(\theta_{aj}) - \sin(\theta_{bj}))^2} \\ &= \sqrt{2(p - \sum_{j=1}^p \cos(\theta_{aj} - \theta_{bj}))} \end{aligned} \quad (36)$$

where p is number of variables. Using the second-order of Taylor series expansion, we have

$$\cos(\theta_{aj} - \theta_{bj}) \cong 1 - (\theta_{aj} - \theta_{bj})^2/2. \quad (37)$$

So, under the assumption that $\theta_{aj} - \theta_{bj}$ is small for all $j \in 1, \dots, p$, the angular distance is

$$d_2(\theta_{a.}, \theta_{b.}) = \sqrt{\sum_{j=1}^p (\theta_{aj} - \theta_{bj})^2} \quad (38)$$

by substituting (37) into (36).

10. CLUSTERING ALGORITHMS

There are various types of clustering algorithms, such as hierarchical clustering algorithm, clustering algorithms based on cost function optimization and etc. In this

section, we review a sub-class of clustering algorithms based on cost function optimization, called hard clustering algorithms, which recover clusters that are as compact as possible. The word ‘hard’ means that each data point is only assigned to a single group. Among these hard clustering algorithms, the k -means and the partitioning around medoids algorithms are mainly discussed.

10.1. Hard clustering algorithm. Suppose that there exists K groups of the underlying X , $\mathbf{C} = \{C_k : 1, 2, \dots, K\}$. Let $\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_K\}$ be K unknown parameter vectors that characterize best the clusters $\mathbf{C} = \{C_k : 1, 2, \dots, K\}$. In a hard clustering algorithm, each data point, \mathbf{x}_i , is assigned to a single group, C_k , exclusively. Let u_{ik} be an indicator function, which has a value 0 or 1. If i -th observation, \mathbf{x}_i , is assigned to the l -th group, C_l , then

$$u_{ik} = \begin{cases} 1 & \text{for } k = l \\ 0 & \text{for all } k \neq l \end{cases} \quad (39)$$

and

$$\sum_{k=1}^K u_{ik} = 1. \quad (40)$$

The cost function is written as

$$J(\Lambda, U) = \sum_{i=1}^n \sum_k^K u_{ik} d(\mathbf{x}_i, \lambda_k), \quad (41)$$

where $d(\mathbf{x}_i, \lambda_k)$ is a dissimilarity measure between the vector \mathbf{x}_i and the cluster representative λ_k .

Let us fix $\lambda_k, k = 1, \dots, K$. $J(\Lambda, U)$ in Equation (41) is minimized if we assign each \mathbf{x}_i to its closest cluster, C_k , that is,

$$u_{ik} = \begin{cases} 1 & \text{if } d(\mathbf{x}_i, \lambda_k) = \min_{l=1,2,\dots,K} d(\mathbf{x}_i, \lambda_l) \\ 0 & \text{otherwise} \end{cases} \quad i = 1, \dots, n. \quad (42)$$

In the case in which two or more minima occur, an arbitrary choice is made. For example, if $d(\mathbf{x}_i, \lambda_k) = d(\mathbf{x}_i, \lambda_l)$ for $k \neq l$, we either take $u_{ik} = 1$ or $u_{il} = 1$.

Now consider the parameter vector λ_k . Taking the gradient of $J(\Lambda, U)$ with respect to λ_k and setting it equal to zero, we obtain

$$\frac{\partial J(\Lambda, U)}{\partial \lambda_k} = \sum_{i=1}^n u_{ik} \frac{\partial d(\mathbf{x}_i, \lambda_k)}{\partial \lambda_k} = \mathbf{0}, \quad k = 1, \dots, K. \quad (43)$$

Let us fix all u_{ik} , and then update $\lambda_k, k = 1, \dots, K$, using (43).

The following iterative algorithm (Theodoridis & Koutroumbas, 2009) is implemented to obtain parameter estimates for U and Λ using Equations (42) and (43).

- set $t = 0$.
- Given the number of groups, K , select $\lambda_k(0)$ randomly as initial values for $\lambda_k, k = 1, \dots, K$.
- Repeat

- For $i = 1, \dots, n$,
 - * For $k = 1, \dots, K$, Update the group membership:¹

$$u_{ik}(t) = \begin{cases} 1 & \text{if } d(\mathbf{x}_i, \lambda_k(t)) = \min_{l=1,2,\dots,K} d(\mathbf{x}_i, \lambda_l(t)) \\ 0 & \text{otherwise,} \end{cases}$$

- * End the loop k .
- End the loop i .
- For $k = 1, \dots, K$,
 - * Update the parameter vectors λ_k by solving

$$\sum_{i=1}^n u_{ik}(t) \frac{\partial d(\mathbf{x}_i, \lambda_k)}{\partial \lambda_k} = \mathbf{0},$$

with respect to λ_k and set $\lambda_k = \lambda_k(t)$ for the current iteration t .

- End the loop k .
 - $t = t + 1$.
- terminate until a termination criterion is met. The termination criterion $\|\Lambda(t) - \Lambda(t-1)\| < \varepsilon$ can be used. Alternatively, the algorithm terminates if U remains unchanged for two successive iterations.

Let the representatives $\lambda_k, k = 1, \dots, K$ be the mean vectors of the clusters $\mathbf{C} = \{C_k : 1, 2, \dots, K\}$. And also, let $d(\mathbf{x}_i, \lambda_k)$ be squared Euclidean distance between the vector \mathbf{x}_i and the mean vector λ_k . For this case, the cost function, (41), becomes

$$J(\Lambda, U) = \sum_{i=1}^n \sum_k^K u_{ik} \|\mathbf{x}_i - \lambda_k\|^2. \quad (44)$$

This is a special case of the generalized hard clustering algorithm known as the k -means algorithm that converges to a minimum of the cost function. The detailed pseudo-code is described in Theodoridis & Koutroumbas (2009, pp 742).

In the k -medoids algorithms, λ_k is a vector that is a representative of the k -th cluster, C_k , for all $k = 1, \dots, K$. This λ_k is called the medoid of the k -th cluster. Let Φ be the set of medoids of all K clusters. I_Φ is the set of indices of the points in X that constitute Φ , whereas $I_{X-\Phi}$ denotes the set of indices of the points that are not medoids. Then, the cost function, for this case, is

$$J(\Lambda, U) = \sum_{i \in I_{X-\Phi}} \sum_{k \in I_\Phi} u_{ik} d(\mathbf{x}_i, \mathbf{x}_k) \quad (45)$$

and

$$u_{ik} = \begin{cases} 1 & \text{if } d(\mathbf{x}_i, \mathbf{x}_k) = \min_{l \in I_\Phi} d(\mathbf{x}_i, \mathbf{x}_l) \\ 0 & \text{otherwise} \end{cases} \quad i = 1, \dots, n \quad (46)$$

¹If two or more minima occur, an arbitrary choice is made.

where $d(\mathbf{x}_i, \mathbf{x}_k)$ is a dissimilarity measure between the points \mathbf{x}_i and \mathbf{x}_k . If the cost function $J(\Lambda, U)$ is minimized, there exists a set of the medoids, Φ , that are the best representatives of the K clusters.

The k -medoids algorithm has two advantages over the k -means algorithm. First, k -medoids algorithms are less sensitive to outliers in comparison to the k -means algorithm. Second, k -medoids algorithms can be applied to data sets originating from either continuous or discrete domains, whereas k -means algorithm can only be used on data sets originating from continuous domains. However, the mean has a geometrical and statistical meaning of the cluster but these properties are not necessarily the case for medoids.

In particular, Kaufman & Rousseeuw (1990) proposed the partitioning around medoids algorithm (or PAM) that minimizes $J(\Lambda, U)$ to determine a set $\Phi \in X$ that are the best representatives of the clusters. The iterative algorithmic scheme, including two stages, is then used to obtain parameter estimates for U and Λ . (a) each \mathbf{x}_i is assigned to a single cluster λ_k using (46), i.e., updating the coefficient u_{ik} for all $i = 1, \dots, n$; (b) If the cost function $J(\Lambda, U)$ can be reduced by interchanging (swapping) an element \mathbf{x}_l in Φ with an element \mathbf{x}_i in $X - \Phi$, then the swap is carried out until the cost function can no longer be decreased. That is, the medoids λ_k are updated in this stage. A detailed procedure can also be found in Theodoridis & Koutroumbas (2009, pp 747).

For each observation i , let $a(i)$ be average dissimilarity between the i -th observation and all other points of the cluster to which the i -th observation belongs. Let the i -th observation belong to the cluster, C_j , where $j \in \{1, \dots, K\}$. For all other $K - 1$ clusters, $C_{k \neq j}$, let $d(\mathbf{x}_i, C_{k \neq j})$ be average dissimilarity of the i -th observation to all observations of $C_{k \neq j}$. The smallest of these $d(\mathbf{x}_i, C_{k \neq j})$ is given as

$$b(i) = \min_{C_{k \neq j}} d(\mathbf{x}_i, C_{k \neq j}),$$

which can be seen as the dissimilarity between the i observation and its nearest cluster to which it does not belong. Then, the silhouette width $s(i)$ is defined in Kaufman & Rousseeuw (1990) as follows:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (47)$$

which gives a value in $[-1, 1]$. If $s(i)$ approaches 1, the i -th observation is probably place in the right cluster. If $s(i)$ takes a value around 0, the observation lies between two clusters. If $s(i)$ is close to -1 , then the i -th observation is probably assigned to the wrong cluster. Further, $\bar{s}(i)$ denotes average silhouette width of all observations, and this quantity assesses quality of clustering as the number of clusters K is given. $\text{ave}_{i \in C_k} s(i)$ is average silhouette width within the cluster k .

11. APPLICATIONS

In this section, the partitioning around medoids (or PAM) algorithm is applied to both the RNA and the protein side-chain data sets. The PAM algorithm is more robust

Group	α	β	γ	δ	ε	ζ
1	5.13	3.06	0.91	1.43	3.62	5.05
2	2.63	3.44	3.06	1.47	3.90	4.92
3	5.04	3.03	0.93	1.46	3.91	2.31

TABLE 3. The medoids of the three clusters are measured in radian.

than the k -means. In particular, the algorithm is flexible to any choice of dissimilarity measure, and is also less sensitive to the outliers. Further, this algorithm uses input as a distance matrix which in our case is based on an angular dissimilarity measure.

For our purpose, we adopt the PAM algorithm implemented in *cluster* package in R programming language, which takes a dissimilarity matrix of all pairs of data points as the input, and gives the group membership label for each data point and the medoid of each cluster as the output. In addition, this implementation also reports the silhouette width for each data point, and the silhouette plot of Kaufman & Rousseeuw (1990) is drawn to assess how good each data is assigned to a group. On the other hand, the dissimilarity matrix is calculated using the Euclidean distance measure $d_2(\theta_a, \theta_b)$, where a, b are any pair of the data. The matrix is implemented in C programming language, and we then call the C code from the PAM function in *cluster* package. In this way, the program allows us to calculate a relatively large data set (up to 10000 data points) for a PC with a 2.0 MHz CPU and a 2G memory installed.

11.1. The RNA data. A random sample of 1000 data points are drawn from the RNA data set. Then, we have a 1000×6 data matrix, X . Each row indicates a data point measured in radian, whereas each column represents a dihedral angle type of $\alpha, \beta, \gamma, \delta, \varepsilon$ and ζ . The dissimilarity matrix with a dimension of 1000×1000 is then calculated based on the data matrix X and the distance measure $d_2(\theta_a, \theta_b)$. After that, cluster analysis is performed on the sample using the PAM algorithm.

We see from figure 11(a) that the PAM algorithm has its maximum average silhouette width around 0.6 at $K = 3$, so the sample has been clustered into 3 groups optimally. Figure 11(b) gives a silhouette plot of the PAM clustering described in Kaufman & Rousseeuw (1990). For each observation i , a bar is drawn, representing its silhouette width $s(i)$. Observations are grouped per cluster, starting from cluster 1 at the top. And also, $s(i)$ are plotted in decreasing order within each cluster. We can see from the figure that these observations are well clustered into 3 groups, although some of them in cluster 2 and 3 have negative silhouette widths. The observations with negative $s(i)$ are probably placed in the wrong cluster.

Figure 12 gives a pairwise scatter plot of the random sample, and each colour labels a group. The group labels are also marked at the medoids that display in Table 3. It can be seen from the figure that there exists the multi-modality feature in cluster 3 in light blue. This is because the PAM algorithm tends to recover a compact collection of data points as a group. So it is sometimes difficult to judge whether these clusters are ‘real’, and how many these ‘real’ clusters exist in the data. Eventually, the experts will

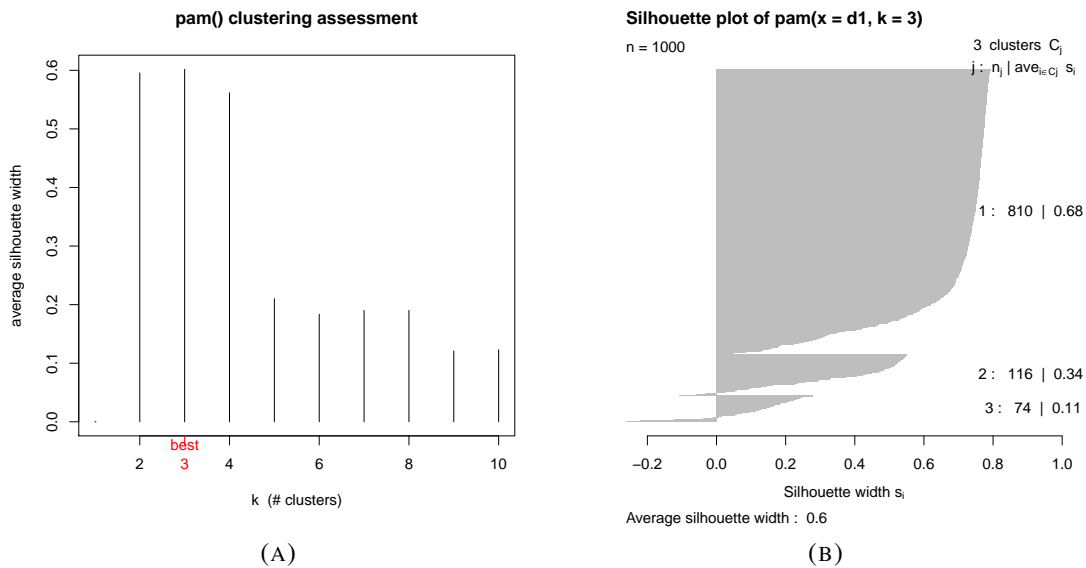


FIGURE 11. (a) gives a plot of average silhouette widths when increasing the number of clusters. The red number will be the ‘best’ number of clusters, which reports the maximum of these average silhouette widths. (b) gives a silhouette plot when the number of clusters equals to 3.

have their own preference to be a cluster and their preferred algorithms. However, in this example, the PAM algorithm picks up the largest group in red consisting of 81% of the total (see Figure 11(b)). The observations that have a secondary structure as α -helix are mainly found in this group.

11.2. The dihedral angles from both the backbone and side-chain of isoleucine.

A random sample of 2000 data points are drawn from the isoleucine data set. Then, we have a 1000×4 data matrix, X . Each row indicates a data point measured in radian, whereas each column represents a dihedral angle type of ϕ , ψ , χ_1 and χ_2 . The dissimilarity matrix with a dimension of 2000×2000 is then calculated based on the data matrix X and the distance measure $d_2(\theta_a, \theta_b)$. After that, cluster analysis is performed on the sample using the PAM algorithm.

We see from Figure 13(a) that the PAM algorithm has its maximum average silhouette width around 0.69 at $K = 8$, so the sample has been clustered into 8 groups optimally. Figure 13(b) gives a silhouette plot of the PAM clustering described in Kaufman & Rousseeuw (1990). For each observation i , a bar is drawn, representing its silhouette width $s(i)$. Observations are grouped per cluster, starting from cluster 1 at the top. We can see from the figure that these observations are well clustered into 8 groups, since the average silhouette width within each group is greater than 0.5 and few observations have negative silhouette widths.

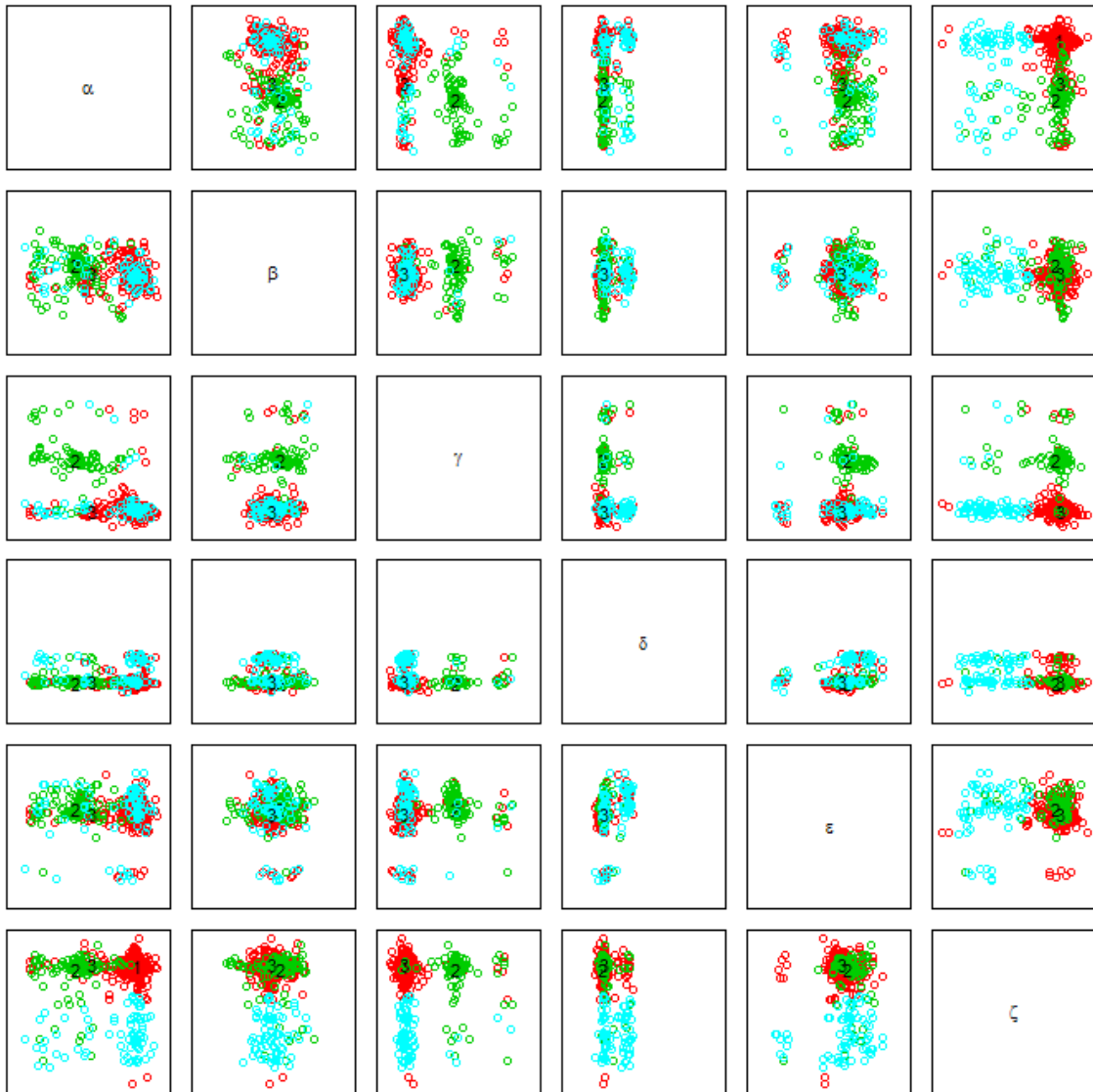


FIGURE 12. A random sample of the RNA data, consisting of 1000 data points, is clustered into 3 groups, each of which is represented by a colour.

Figure 14 gives a pairwise scatter plot of the random sample. To visualize the group memberships, observations of the odd group labels are in red whereas observations of the even group labels are in green. And also, the group labels are marked at the medoids that display in Table 4. It can be seen from the figure that there are some potential ‘groups’ that may be identified by human eyes, but they are not considered to be clusters by the PAM algorithm when $K = 8$. For example, on the scatter plot of χ_1 against χ_2 , the observations occurring at the left and right bottoms are not considered

Group	ϕ	ψ	χ_1	χ_2
1	1.00	5.90	1.09	2.98
2	2.03	2.38	5.10	2.93
3	1.23	5.33	5.23	2.98
4	1.32	5.32	5.34	5.25
5	1.99	2.36	5.15	5.27
6	1.99	2.67	3.38	1.12
7	1.41	3.08	1.09	2.95
8	0.89	5.60	3.27	2.91

TABLE 4. The medoids of the three clusters are measured in radian.

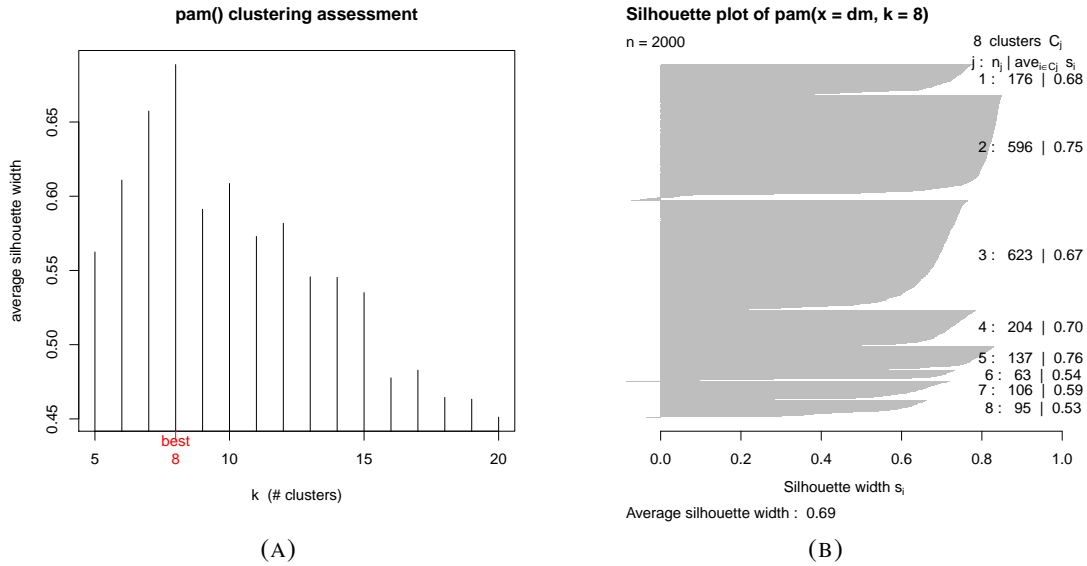


FIGURE 13. (a) gives a plot of average silhouette widths when increasing the number of clusters. The red number will be the ‘best’ number of clusters, which reports the maximum of these average silhouette widths. (b) gives a silhouette plot when the number of clusters equals to 8.

to be two individual clusters. This is because only few observations are very close together, and the PAM algorithm may treat these observations as outliers when $K = 8$.

In the context of EM for mixtures, the clusters found by the PAM algorithm can be used to produce the initial values for the parameters to be estimated. In particular, their medoids are taken as initial values for the mean parameters of the mixture model. The initial values for the mixing proportions are calculated by counting the number of observations within each cluster found by the PAM algorithm. Further, other initial values can also be computed based on the group membership of each observation.

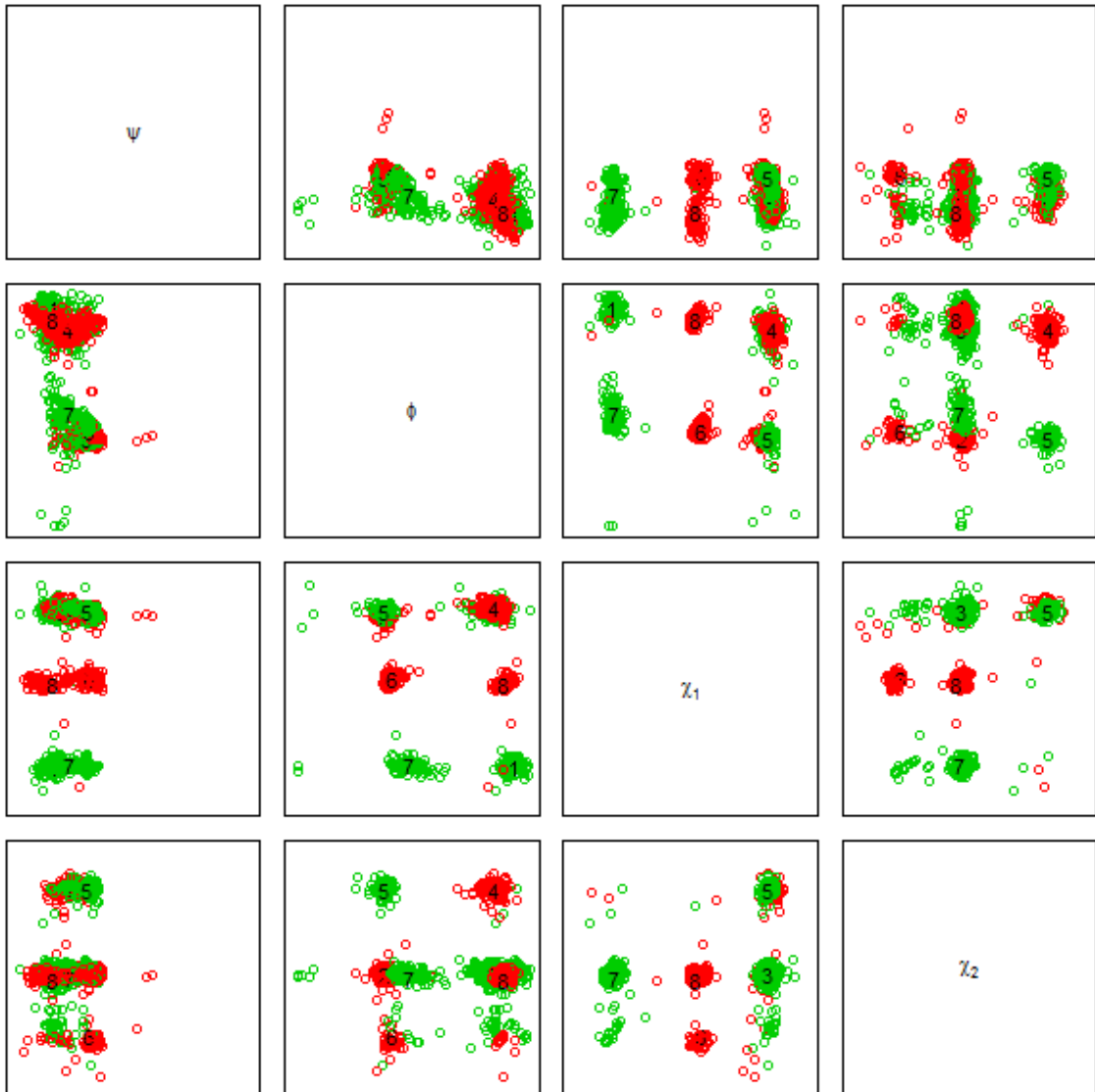


FIGURE 14. A random sample of the protein side data, consisting of 2000 data points, is clustered into 8 groups. Red represents any even label number, while there is green for the odd label numbers.

12. CONCLUSIONS

First of all, we reviewed the circular correlation coefficient given in Jammalamadaka *et al.* (2001), and found that this correlation coefficient works well even when data is semi-concentrated. After that, we introduced several circular correlation matrix functions, based on circular correlation coefficient as defined in (3), to investigate correlation of the six RNA backbone dihedral angles. The sample correlation matrix, $\hat{\Phi}(k)$, is calculated with lag $k = 1, \dots, 5$, and it reports large values when the lag $k =$

1,2. Finally, we found that $\hat{\Psi}$ is preferable to $\hat{\Phi}(0)$, since $\hat{\Psi}$ captures the relationship of these dihedral angles around a circle.

Secondly, we discussed the existing circular PCA methods and its properties, summarized in Table 5(a). We then applied these method on a variety of the data sets simulated from both the sine and cosine distribution. The principal components are calculated for each data set for each method, shown in Table 6. Further, the scores are plotted on the first a few principal components. Table 5(b) gives a summary of visualizations based on the simulated data sets. Finally, we discussed the strength and weakness of each method in Table 7.

All the methods are applicable to the concentrated data, but become problematic when data become diffuse (or semi-diffuse). For moderate data, the angular method may give a reasonable visualization and explanations in terms of eigenvalues and eigenvectors. For diffuse data, all the methods fail to give some reasonable visualizations. For semi-concentrated data, angular, sine and complex methods may be useful.

Recently, Mu *et al.* (2005) and Altis *et al.* (2007) applied this Euclidean methods to the dihedral angles coming from molecular dynamics simulations. They claimed the method reduces dimensionality of original angle distribution and identifies groups on the first few PCs. It means that the method could be used in clustering analysis. However, from our study, the method would not be useful for diffuse data (and some moderate data). The method can only be used if data is concentrated.

Finally, we introduced the two dissimilarity measures between any pair of data points on the p -fold torus. Cluster analysis is then performed on both the RNA and isoleucine data sets using the PAM algorithm and one of these dissimilarity measures. Given the number of groups K , the PAM algorithm assigns each data point into one of these K clusters. In the context of EM algorithm for mixtures, these clusters can be used to produce the initial values for the parameters to be estimated.

ACKNOWLEDGEMENT

We are grateful to Mark Gilson for drawing our attention to the problem of circular PCA. We also would like to thanks Thomas Hamelryck and his group for the RNA and protein data. This work is funded by a Dorothy Hodgkin postgraduate award co-sponsored between BBSRC and GlaxoSmithKline.

REFERENCES

- ALTIS, A., NGUYEN, P., HEGGER, R. & STOCK, G. (2007). Dihedral angle principal component analysis of molecular dynamics simulations. *The Journal of Chemical Physics*, **126**, 244111.
- FRELLSEN, J., MOLTKE, I., THIIM, M., MARDIA, K., FERKINGHOFF-BORG, J. & HAMELRYCK, T. (2009). A probabilistic model of RNA conformational space. *PLoS Comput Biol*, **5**, e1000406.
- JAMMALAMADAKA, S. & SARMA, Y. (1988). *A correlation coefficient for angular variable. Statistical Theory and Data Analysis 2*. North Holland.

- JAMMALAMADAKA, S., RAO, S. & SENGUPTA, A. (2001). *Topics in Circular Statistics*. World Scientific Press.
- KAUFMAN, L. & ROUSSEEUW, P. (1990). *Finding Groups in Data*. John Wiley & Sons.
- MARDIA, K. & JUPP, P. (1999). *Directional Statistics*. WileyBlackwell.
- MARDIA, K., KENT, J. & BIBBY, J. (1979). *Multivariate Analysis*. Academic Press.
- MARDIA, K., TAYLOR, C. & SUBRAMANIAM, G. (2007). Protein bioinformatics and mixtures of bivariate von mises distributions for angular data. *Biometrics*, **63**, 505–512.
- MU, Y., NGUYEN, P.H. & STOCK, G. (2005). Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *PROTEINS: Structure, Function, and Bioinformatics*, **58**, 45–52.
- MURRAY, L., III, W.A., RICHARDSON, D. & RICHARDSON, J. (2003). RNA backbone is rotameric. *PNAS*, **100**, 1390413909.
- THEODORIDIS, S. & KOUTROUMBAS, K. (2009). *Pattern Recognition*. Academic Press.
- WEI, W.W. (1989). *Time series analysis*. Addison-Wesley.

APPENDIX A. THE CIRCULAR CORRELATION MATRIX FUNCTIONS AND THEIR P-VALUE MATRICES

The circular correlation matrix function of 6 dihedral angles with lag 1 is

$$\hat{\Phi}(1) = \begin{pmatrix} & \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \\ \alpha & -0.023 & 0.006 & -0.001 & 0.016 & 0.027 & -0.012 \\ \beta & -0.009 & 0.035 & 0.040 & 0.026 & 0.012 & 0.009 \\ \gamma & 0.020 & -0.001 & -0.010 & 0.004 & -0.009 & -0.007 \\ \delta & 0.014 & 0.079 & 0.062 & 0.303 & 0.192 & -0.031 \\ \varepsilon & 0.151 & -0.029 & -0.018 & 0.222 & 0.157 & -0.040 \\ \zeta & -0.154 & -0.023 & 0.005 & -0.110 & -0.098 & -0.028 \end{pmatrix}$$

and its P-values

$$\begin{pmatrix} & \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \\ \alpha & 0.0793 & 0.6349 & 0.9590 & 0.2154 & 0.0292 & 0.3065 \\ \beta & 0.4879 & 0.0125 & 0.0025 & 0.0362 & 0.3116 & 0.4270 \\ \gamma & 0.1118 & 0.9547 & 0.4353 & 0.7601 & 0.4369 & 0.5690 \\ \delta & 0.3673 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0223 \\ \varepsilon & 0.0000 & 0.0679 & 0.2420 & 0.0000 & 0.0000 & 0.0023 \\ \zeta & 0.0000 & 0.1822 & 0.7147 & 0.0000 & 0.0000 & 0.0477 \end{pmatrix}.$$

The circular correlation matrix function of 6 dihedral angles with lag 2 is

$$\hat{\Phi}(2) = \begin{pmatrix} & \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \\ \alpha & 0.016 & -0.012 & -0.030 & 0.013 & 0.000 & -0.004 \\ \beta & -0.026 & 0.037 & 0.019 & 0.028 & 0.009 & -0.011 \\ \gamma & -0.012 & 0.020 & 0.035 & 0.002 & 0.016 & -0.011 \\ \delta & 0.009 & 0.056 & 0.006 & 0.115 & 0.081 & 0.009 \\ \varepsilon & 0.001 & 0.012 & -0.004 & 0.108 & 0.086 & -0.007 \\ \zeta & -0.038 & 0.016 & -0.013 & -0.036 & -0.033 & -0.028 \end{pmatrix}$$

and its P-values

$$\begin{pmatrix} & \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \\ \alpha & 0.1803 & 0.3131 & 0.0126 & 0.2696 & 0.9824 & 0.7420 \\ \beta & 0.0254 & 0.0025 & 0.1089 & 0.0164 & 0.4540 & 0.3493 \\ \gamma & 0.2939 & 0.1071 & 0.0036 & 0.8217 & 0.1475 & 0.3804 \\ \delta & 0.4601 & 0.0000 & 0.5990 & 0.0000 & 0.0000 & 0.4137 \\ \varepsilon & 0.9281 & 0.3503 & 0.7565 & 0.0000 & 0.0000 & 0.5267 \\ \zeta & 0.0033 & 0.2527 & 0.3036 & 0.0074 & 0.0080 & 0.0192 \end{pmatrix}.$$

The circular correlation matrix function of 6 dihedral angles with lag 3 is

$$\hat{\Phi}(3) = \begin{pmatrix} & \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \\ \alpha & 0.014 & -0.009 & -0.026 & -0.003 & 0.002 & 0.001 \\ \beta & -0.004 & 0.031 & -0.016 & 0.005 & 0.011 & -0.020 \\ \gamma & 0.000 & 0.010 & 0.025 & 0.004 & 0.018 & -0.011 \\ \delta & 0.008 & 0.025 & -0.021 & 0.026 & 0.007 & -0.003 \\ \varepsilon & 0.016 & 0.029 & -0.033 & 0.037 & 0.031 & -0.014 \\ \zeta & 0.024 & -0.033 & -0.010 & -0.010 & 0.010 & 0.004 \end{pmatrix}$$

and its P-values

$$\begin{pmatrix} & \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \\ \alpha & 0.2115 & 0.4254 & 0.0265 & 0.7711 & 0.8775 & 0.9258 \\ \beta & 0.6949 & 0.0085 & 0.1699 & 0.6329 & 0.3383 & 0.0587 \\ \gamma & 0.9924 & 0.3967 & 0.0311 & 0.7338 & 0.1060 & 0.3499 \\ \delta & 0.4799 & 0.0192 & 0.0389 & 0.0308 & 0.5299 & 0.7904 \\ \varepsilon & 0.1605 & 0.0109 & 0.0026 & 0.0011 & 0.0042 & 0.2249 \\ \zeta & 0.0381 & 0.0054 & 0.3949 & 0.3768 & 0.3380 & 0.7239 \end{pmatrix}.$$

The circular correlation matrix function of 6 dihedral angles with lag 4 is

$$\hat{\Phi}(3) = \begin{pmatrix} & \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \\ \alpha & 0.008 & -0.030 & -0.009 & 0.017 & 0.023 & -0.003 \\ \beta & -0.002 & 0.023 & -0.002 & -0.006 & -0.012 & 0.008 \\ \gamma & 0.010 & 0.019 & -0.006 & 0.006 & -0.007 & -0.009 \\ \delta & 0.003 & -0.005 & -0.016 & -0.005 & -0.008 & 0.005 \\ \varepsilon & 0.024 & -0.013 & 0.003 & 0.001 & 0.009 & 0.007 \\ \zeta & -0.014 & -0.005 & 0.008 & 0.016 & 0.018 & 0.028 \end{pmatrix}$$

and its P-values

$$\begin{pmatrix} & \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \\ \alpha & 0.4618 & 0.0056 & 0.3885 & 0.1077 & 0.0364 & 0.7954 \\ \beta & 0.8309 & 0.0407 & 0.8200 & 0.5707 & 0.2595 & 0.4922 \\ \gamma & 0.3811 & 0.0956 & 0.5534 & 0.5857 & 0.5673 & 0.4336 \\ \delta & 0.8010 & 0.6476 & 0.1289 & 0.6753 & 0.4678 & 0.6583 \\ \varepsilon & 0.0302 & 0.2221 & 0.7627 & 0.9267 & 0.3903 & 0.5295 \\ \zeta & 0.1951 & 0.6516 & 0.4981 & 0.1235 & 0.1112 & 0.0107 \end{pmatrix}.$$

The circular correlation matrix function of 6 dihedral angles with lag 5 is

$$\hat{\Phi}(3) = \begin{pmatrix} & \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \\ \alpha & 0.008 & -0.004 & -0.005 & 0.019 & 0.014 & 0.009 \\ \beta & -0.014 & 0.004 & 0.017 & -0.019 & -0.012 & 0.001 \\ \gamma & -0.039 & -0.001 & 0.040 & -0.019 & 0.003 & 0.009 \\ \delta & 0.016 & -0.006 & -0.027 & -0.019 & -0.007 & -0.001 \\ \varepsilon & 0.009 & 0.011 & -0.003 & -0.020 & -0.005 & 0.003 \\ \zeta & 0.009 & 0.028 & -0.005 & 0.013 & 0.003 & -0.017 \end{pmatrix}$$

and its P-values

$$\begin{pmatrix} & \alpha & \beta & \gamma & \delta & \varepsilon & \zeta \\ \alpha & 0.4808 & 0.6860 & 0.6255 & 0.0692 & 0.2226 & 0.4305 \\ \beta & 0.2260 & 0.7152 & 0.1317 & 0.0685 & 0.2869 & 0.8956 \\ \gamma & 0.0010 & 0.9161 & 0.0009 & 0.0583 & 0.7609 & 0.4457 \\ \delta & 0.1136 & 0.5687 & 0.0063 & 0.0869 & 0.5308 & 0.9132 \\ \varepsilon & 0.3942 & 0.3220 & 0.7826 & 0.0646 & 0.6650 & 0.8164 \\ \zeta & 0.3930 & 0.0123 & 0.6430 & 0.2084 & 0.7568 & 0.1032 \end{pmatrix}.$$

APPENDIX B. SOME PROPERTIES OF THE EUCLIDEAN METHOD

Let (θ, ϕ) be a pair of angular random variables with the mean direction (μ, ν) . Consider the Euclidean method with the mean direction centered at $\mathbf{0}$, then the first principal component can be written as:

$$\begin{aligned} l_1(\mu, \nu) &= a_1 \sin(\theta - \mu) + a_2 \cos(\theta - \mu) + a_3 \sin(\phi - \nu) + a_4 \cos(\phi - \nu) \\ &= a_1 \sin(\theta) \cos(\mu) - a_1 \cos(\theta) \sin(\mu) + a_2 \cos(\theta) \cos(\mu) + a_2 \sin(\theta) \sin(\mu) \\ &\quad + a_3 \sin(\phi) \cos(\nu) - a_3 \cos(\phi) \sin(\nu) + a_4 \cos(\phi) \cos(\nu) + a_4 \sin(\phi) \sin(\nu) \\ &= \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} \cos(\mu) & -\sin(\mu) \\ \sin(\mu) & \cos(\mu) \end{pmatrix} \begin{pmatrix} \sin(\theta) \\ \cos(\theta) \end{pmatrix} \\ &\quad + \begin{pmatrix} a_3 & a_4 \end{pmatrix} \begin{pmatrix} \cos(\nu) & -\sin(\nu) \\ \sin(\nu) & \cos(\nu) \end{pmatrix} \begin{pmatrix} \sin(\phi) \\ \cos(\phi) \end{pmatrix} \\ \text{Let } \begin{pmatrix} a'_1 & a'_2 \end{pmatrix} &= \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} \cos(\mu) & -\sin(\mu) \\ \sin(\mu) & \cos(\mu) \end{pmatrix}, \\ \text{and let } \begin{pmatrix} a'_3 & a'_4 \end{pmatrix} &= \begin{pmatrix} a_3 & a_4 \end{pmatrix} \begin{pmatrix} \cos(\nu) & -\sin(\nu) \\ \sin(\nu) & \cos(\nu) \end{pmatrix}. \\ &= a'_1 \sin(\theta) + a'_2 \cos(\theta) + a'_3 \sin(\phi) + a'_4 \cos(\phi) \\ &= l_1^*(\mu = 0, \nu = 0) \end{aligned}$$

Clearly, there is a link between l_1 and l_1^* , by a rotation, mathematically,

$$\mathbb{A} = \begin{pmatrix} \mathbf{A}(\mu) & \mathbf{0} \\ \mathbf{0} & \mathbf{A}(\nu) \end{pmatrix}$$

where $\mathbf{A}(\cdot) = \begin{pmatrix} \cos(\cdot) & -\sin(\cdot) \\ \sin(\cdot) & \cos(\cdot) \end{pmatrix}$. Thus, $l_1 = \mathbb{A}l_1^*$.

In conclusion, Euclidean circular PCA is invariant if the mean direction is centered at $\mathbf{0}$ or not. Note that any circular PCA method is invariant under changing sign(s) of eigenvector(s) of the covariance matrix.

APPENDIX C. COMPUTE THE MOMENT $\mathbb{E}(\cos(\theta) \sin(\phi))$

If $f_s(\theta, \phi)$ is a density function of the sine model, then $\mathbb{E}(\cos(\theta) \sin(\phi)) = 0$.

Proof.

$$\begin{aligned}
& \mathbb{E}(\cos(\theta) \sin(\phi)) \\
&= \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \cos(\theta) \sin(\phi) \exp\{k_1 \cos(\theta) + k_2 \cos(\phi) + \lambda \sin(\theta) \sin(\phi)\} d\theta d\phi \\
&= \int_0^{\pi} \int_0^{\pi} \cos(\theta) \sin(\phi) \exp\{k_1 \cos(\theta) + k_2 \cos(\phi) + \lambda \sin(\theta) \sin(\phi)\} d\theta d\phi \\
&+ \int_0^{\pi} \int_{-\pi}^0 \cos(\theta) \sin(\phi) \exp\{k_1 \cos(\theta) + k_2 \cos(\phi) + \lambda \sin(\theta) \sin(\phi)\} d\theta d\phi \\
&+ \int_{-\pi}^0 \int_0^{\pi} \cos(\theta) \sin(\phi) \exp\{k_1 \cos(\theta) + k_2 \cos(\phi) + \lambda \sin(\theta) \sin(\phi)\} d\theta d\phi \\
&+ \int_{-\pi}^0 \int_{-\pi}^0 \cos(\theta) \sin(\phi) \exp\{k_1 \cos(\theta) + k_2 \cos(\phi) + \lambda \sin(\theta) \sin(\phi)\} d\theta d\phi,
\end{aligned}$$

since $f_s(\theta, \phi)$ is reflection symmetric about x and y axis. If change the variable θ to $-\theta$, then

$$\begin{aligned}
& \int_0^{\pi} \int_{-\pi}^0 \cos(\theta) \sin(\phi) \exp\{k_1 \cos(\theta) + k_2 \cos(\phi) + \lambda \sin(\theta) \sin(\phi)\} d\theta d\phi \\
&= \int_0^{\pi} \int_0^{\pi} -\cos(\theta) \sin(\phi) \exp\{k_1 \cos(\theta) + k_2 \cos(\phi) - \lambda \sin(\theta) \sin(\phi)\} d\theta d\phi;
\end{aligned}$$

If change the variable ϕ to $-\phi$, then

$$\begin{aligned}
& \int_{-\pi}^0 \int_0^{\pi} \cos(\theta) \sin(\phi) \exp\{k_1 \cos(\theta) + k_2 \cos(\phi) + \lambda \sin(\theta) \sin(\phi)\} d\theta d\phi \\
&= \int_0^{\pi} \int_0^{\pi} \cos(\theta) \sin(\phi) \exp\{k_1 \cos(\theta) + k_2 \cos(\phi) - \lambda \sin(\theta) \sin(\phi)\} d\theta d\phi
\end{aligned}$$

and if change both (θ, ϕ) to $(-\theta, -\phi)$, then

$$\begin{aligned}
& \int_{-\pi}^0 \int_{-\pi}^0 \cos(\theta) \sin(\phi) \exp\{k_1 \cos(\theta) + k_2 \cos(\phi) + \lambda \sin(\theta) \sin(\phi)\} d\theta d\phi \\
&= \int_0^{\pi} \int_0^{\pi} -\cos(\theta) \sin(\phi) \exp\{k_1 \cos(\theta) + k_2 \cos(\phi) + \lambda \sin(\theta) \sin(\phi)\} d\theta d\phi.
\end{aligned}$$

Thus, sum of the four separate integrals is

$$\mathbb{E}(\cos(\theta) \sin(\phi)) = 0.$$

□

TABLE 5. (a) a description of the four Circular PCA methods, and (b) a summary of visualizations based on our simulated data sets.

(A)

	Transformation	Properties
Angular circular PCA	$x_j^{[1]} = [(\theta_j - \mu_j - \pi) \bmod 2\pi] - \pi$	Invariant under sign changed
Sine circular PCA	$x_j^{[2]} = \sin(\theta_j - \mu_j)$	Invariant under sign changed
Euclidean circular PCA	$\mathbf{x}_j^{[3]} = (\sin(\theta_j), \cos(\theta_j))$ (Altis <i>et al.</i> , 2007)	Invariant under sign changed
Complex circular PCA	$\mathbf{x}_j^{[4]} = \exp\{i(\theta_j - \mu_j)\}$ (Altis <i>et al.</i> , 2007)	Invariant under rotation

(B)

	Concentrated	Moderate	Diffuse	Semi-concentrated
Ang.	Rotate by $\pi/4$	Rotate by $\pi/4$	Rotate by $\pi/4$	Rotate by $\pi/2$
Sin	Similar to Ang.	Diamond shape (folding)	Diamond shape	$\sin(\phi)$ against $\sin(\theta)$
Euc.	Similar to Ang.	Diamond shape	Full filled Circle	Wrap ϕ around a circle
Arg of Cpx.	Two lines on torus	Two lines on torus	Two lines on torus	Rotate by $\pi/2$
Mod of Cpx.	Positive part of a circle	Positive part of a circle	Positive part of a circle	Straight line

TABLE 6. A Summary of the circular PCA methods in terms of performance

	Concentrated	Moderate	Diffuse	Semi-concentrated
Ang.	$\theta + \phi$ and $\phi - \theta$ 1 st PC is dominated.	Same 2 nd eigenvalue counts a proportion of the total.	Same Both two PCs are important.	Approx. ϕ and θ 2 nd eigenvalue counts a proportion of the total.
Sin	Approx. $\theta + \phi$ and $\phi - \theta$ 1 st PC is dominated.	$\sin(\theta) + \sin(\phi)$ and $\sin(\phi) - \sin(\theta)$ 2 nd eigenvalue counts a proportion of the total.	$\sin(\theta) + \sin(\phi)$ and $\sin(\phi) - \sin(\theta)$ Both two PCs are important.	Approx. $\sin(\phi)$ and $\sin(\theta)$ 2 nd eigenvalue counts a proportion of the total.
Euc.	Approx. $\theta + \phi$ and $\phi - \theta$ 1 st PC is dominated.	$\sin(\theta) + \sin(\phi)$ and $\sin(\phi) - \sin(\theta)$ The first two PCs are important.	$\sin(\theta) + \sin(\phi)$, $\cos(\phi) + \cos(\theta)$ $\cos(\theta) - \cos(\phi)$ and $\sin(\phi) - \sin(\theta)$ All four PCs are important.	Approx. $\sin(\phi)$ and $\cos(\phi)$ The first two PCs are important than the others.
Cpx.	Approx. $\theta + \phi$ and $\phi - \theta$ 1 st PC is dominated.	$\exp(i\theta) + \exp(i\phi)$ and $\exp(i\theta) - \exp(i\phi)$ 2 nd eigenvalue counts a proportion of the total.	$\exp(i\theta) + \exp(i\phi)$ and $\exp(i\theta) - \exp(i\phi)$ Both two PCs are important.	Approx. $\exp(i\phi)$ and $\exp(i\theta)$ 2 nd eigenvalue counts a proportion of the total.

TABLE 7. Strength&weakness

	Strength	Weakness
The Angular Circular PCA	Simple and straightforward.	Illegitimate to treat angles as numbers. It matters when data are diffuse, e.g., $(179^\circ, 0)$ and $(-179^\circ, 0)$.
The Sine Circular PCA	Map angles to some number between -2 and 2.	Not involve angles any more. The wrapping problem due to nature of sine function.
The Euclidean Circular PCA	The principal components are written in terms of some linear combinations of sine and cosine, and ordered by magnitudes of the corresponding eigenvalues.	The number of principal components is doubled, i.e., length of $2p$.
The Complex Circular PCA	Polar representation of the complex principal components.	Treat the sine terms as equally important as the cosine terms (cf: the Euclidean CPCA).