Discrete Circular Distributions with Applications to Shared Orthologs of Paired Circular Genomes

Tomoaki Imoto[1, *], Grace S. Shieh[2, *] and Kunio Shimizu[3]

Abstract: For structural comparisons of paired prokaryotic genomes, an important topic in synthetic and evolutionary biology, the locations of shared orthologous genes (henceforth orthologs) are observed as binned data. This and other data, e.g., wind directions recorded at monitoring sites and intensive care unit arrival times on the 24-hour clock, are counted in binned circular arcs, thus modeling them by discrete circular distributions (DCDs) is required. We propose a novel method to construct a DCD from a base continuous circular distribution (CCD). The probability mass function is defined to take the normalized values of the probability density function at some pre-fixed equidistant points on the circle. Five families of constructed DCDs which have normalizing constants in closed form are presented. Simulation studies show that DCDs outperform the corresponding CCDs in modeling grouped (discrete) circular data, and minimum chi-square estimation outperforms maximum likelihood estimation for parameters. We apply the constructed DCDs, invariant wrapped Poisson and wrapped discrete skew Laplace to compare the structures of paired bacterial genomes. Specifically, discrete four-parameter wrapped Cauchy (nonnegative trigonometric sums) distribution models multi-modal shared orthologs in *Clostridium* (*Sulfolobus*) better than the others considered, in terms of AIC and Freedman's goodness-of-fit test. The result that different DCDs fit the shared orthologs is consistent with the fact they belong to two kingdoms. Nevertheless, these prokaryotes have a common favored site around 70° on the unit circle; this finding is important for building synthetic prokaryotic genomes in synthetic biology. These DCDs can also be applied to other binned circular data.

Keywords: Bacterial genomes, circular distribution, goodness-of-fit test, modeling, synthetic and evolutionary biology.

[1] School of Management and Information, University of Shizuoka, Shizuoka, Japan.

[2] Institute of Statistical Science, Academia Sinica, Taipei, Taiwan.

[3] School of Statistical Thinking, The Institute of Statistical Mathematics, Tokyo, Japan.

* Corresponding Authors: Tomoaki Imoto. Email: imoto0923@gmail.com;
Grace S. Shieh. Email: gshieh@stat.sinica.edu.tw.

Received: 29 August 2019; Accepted: 07 November 2019.

CMES. doi:10.32604/cmes.2020.08466 www.techscience.com/journal/CMES

1 Introduction

Most prokaryotic genomes (1158 out of 1194, NCBI, August 2010) are made up of single circular chromosome (henceforth called circular genomes). Here, our emphasis is on the structure of circular genomes, which plays an important role in synthetic and evolutionary biology. For example, the most and least favored regions in which shared orthologous genes (henceforth orthologs) between bacterial genomes are located are of interest and important; orthologs are genes directly evolved from an ancestoral gene [Tatusov, Koonin and Lipman (1997)], and can be traced through different species across evolution. Further, genome organization may influence gene expression, which is vital for organisms [Carrera, Rodrigo and Jaramillo (2009)].

While studying structural comparisons between paired circular genomes [Shieh, Zheng, Johnson et al. (2011)], we found that the locations of shared orthologs are often observed as binned data. Since shared orthologs and other data, e.g., wind directions recorded at monitoring sites and intensive care unit arrival times on the 24-hour clock, are counted in binned circular arcs, modeling them by discrete circular distributions (DCDs) is required, which was the motivation for this study.

Circular (angular) data can be represented as points on the circumference of a unit circle, e.g., wind directions at a monitoring site, and others [Fisher (1993); Jammalamadaka and SenGupta (2001); Johnson and Wehrly (1977); Mardia and Jupp (2000)]. Circular data are modeled by distributions on the circle, namely circular distributions. Continuous circular distributions (CCDs) can be generated using several methods such as projection and wrapping. The wrapping approach is an effective method for generating a probability density function (pdf) on the circle from a pdf on the line. For example, a wrapped normal distribution is constructed by wrapping a normal distribution onto the unit circle. Similarly, any distribution on integers can be wrapped around the circumference of a unit circle to construct the probability mass function (pmf) on the circle. For instance, the wrapped Poisson distribution is constructed by wrapping Poisson distribution onto the unit circle [Mardia and Jupp (2000)]. Moreover, wrapped discrete skew Laplace (WDSL) and wrapped geometric distributions have been studied [Jayakumar and Jacob (2012); Jacob and Jayakumar (2013)], respectively. Recently, Mastrantonio and colleagues introduced the invariant wrapped Poisson (IWP) distribution and investigated the invariance properties [Mastrantonio, Jona Lasinio, Maruotti et al. (2015)]. The wrapping method can immediately constuct the corresponding circular distribution from a discrete distribution on R^1; however, the normalizing constants are not analytic in general.

In this article, we propose a novel method to construct DCDs for discrete circular data. The constructed DCD is defined on the n (≥ 2) pre-fixed equidistant points $2\pi j/n$ (j=0, 1, 2, …, $n-1$) of the circumference of a unit circle from the pdf $f(\theta)$ of the corresponding CCD on $[0, 2\pi)$. The pmf of the DCD from a base pdf $f(\theta)$ is defined by

$$p_n(j) = C_n^{-1} f\left(\frac{2\pi j}{n}\right), j = 0, 1, 2, \ldots, n-1 \tag{1}$$

with the normalizing constant $C_n = \sum_{j=0}^{n-1} f\left(\frac{2\pi j}{n}\right)$. We note that in general, the number of divisions is determined by domain knowledge. For example, n=24 is commonly used for hourly data such as gun crime data [Mastrantonio, Jona Lasinio, Maruotti et al. (2015)] and intensive care unit arrival times on the 24-hour clock, while n=8 and 16 are used for wind directions at a monitoring site. As $n \to \infty$, the explicit form of $p_n(j)$ is shown in Appendix A.

In the following section, we derive the normalizing constant of Eq. (1), then we present the characteristic function and trigonometric moments of DCDs. Next, families of discrete distributions constructed from von Mises, cardioid, nonnegative trigonometric sums [Fernández-Durán (2004)], wrapped Cauchy and four-parameter wrapped Cauchy [Kato and Jones (2015)] distributions are illustrated, and their trigonometric moments are presented. Furthermore, we study how well DCDs and the corresponding CCDs model the grouped circular data, and compares two estimation methods via Monte Carlo simulations. Finally, we apply the constructed DCDs, IWP and WDSL to model marginals of the shared orthologs between a pair of circular genomes; these prokaryotes belong to different kingdoms. The fitting results of our DCDs, IWP and WDSL are also compared. We close with some discussion in Section 6.

2 Methods

In this section, we derive the expressions of the normalizing constant and trigonometric moments of the pmf in Eq. (1); the former is required by a DCD while the latter are basic properties. For this purpose, the following proposition is useful.

Proposition 1. *Let $i = \sqrt{-1}$ denote the imaginary unit. When k is an integer such that $0 \le k \le n - 1$, then we have the following.*

1. For $p = \ldots, -3n, -2n, -n, 0, n, 2n, 3n, \ldots$, it holds that

$$\sum_{j=0}^{k} e^{i(2\pi pj/n)} = k + 1,$$

and thus,

$$\sum_{j=0}^{k} \cos\left(\frac{2\pi pj}{n}\right) = k + 1, \quad \sum_{j=0}^{k} \sin\left(\frac{2\pi pj}{n}\right) = 0.$$

2. For an integer p such that $p \neq \ldots, -3n, -2n, -n, 0, n, 2n, 3n \ldots$ *, it holds that*

$$\sum_{j=0}^{k} e^{i(2\pi pj/n)} = \frac{\sin\{\pi p(k+1)/n\}}{\sin(\pi p/n)} e^{i(\pi pk/n)},$$

and thus, for k=n−1,

$$\sum_{j=0}^{n-1} \cos\left(\frac{2\pi pj}{n}\right) = \sum_{j=0}^{n-1} \sin\left(\frac{2\pi pj}{n}\right) = 0.$$

The first part of Proposition 1 is obvious and the second part is proved by induction. Hereafter, we suppose that $f(\theta)$ is represented by a Fourier series

$$f(\theta) = \frac{1}{2\pi} \sum_{p=-\infty}^{\infty} \phi_p e^{-ip\theta} = \frac{1}{2\pi} \left[1 + 2\sum_{p=1}^{\infty} \{\alpha_p \cos(p\theta) + \beta_p \sin(p\theta)\} \right],$$

where ϕ_p, α_p and β_p represent the characteristic function, pth cosine moment and pth sine moment of a circular random variable Θ with pdf $f(\theta)$ respectively, i.e., ϕ_p=E[$e^{ip\Theta}$], α_p=E[cos(pΘ)] and β_p=E[sin(pΘ)].

2.1 The formula for the normalizing constants

From Proposition 1 and the Fourier series expansion for $f(\theta)$, we find the equation, for k=0, 1, 2, …, n−1,

$$\sum_{j=0}^{k} f\left(\frac{2\pi j}{n}\right) = \frac{1}{2\pi} \left[(k+1)\left(1 + 2\sum_{p=1}^{\infty} \alpha_{np}\right) + 2 \times \sum_{\substack{p=1,2,3,\ldots,\\ p\neq n,2n,3n,\ldots}} \frac{\sin\{\pi p(k+1)/n\}}{\sin(\pi p/n)} \times \left\{ \alpha_p \cos\left(\frac{\pi pk}{n}\right) + \beta_p \sin\left(\frac{\pi pk}{n}\right) \right\} \right].$$

For $k = n - 1$ in the above equation, it holds that

$$\sum_{j=0}^{n-1} f\left(\frac{2\pi j}{n}\right) = \frac{n}{2\pi} \left(1 + 2\sum_{p=1}^{\infty} \alpha_{np}\right).$$

This equation leads to the identity of the normalizing constant C_n in (1) as

$$C_n = \frac{n}{2\pi} \left(1 + 2\sum_{k=1}^{\infty} \alpha_{nk}\right).$$

2.2 Trigonometric moments

Suppose that a random variable Θ_n follows a DCD with the pmf (1). To obtain its characteristic function $\phi_{n,q}$=E[$e^{iq\Theta_n}$]=$\alpha_{n,q}$+$i\beta_{n,q}$, it suffices to calculate the cases q=0, 1, 2, …, n−1, because obviously, $\phi_{n,q+kn}$=$\phi_{n,q}$ for k=$1, 2, 3, \ldots$, and $\alpha_{n,-q}$=$\alpha_{n,q}$ and $\beta_{n,-q}$= −$\beta_{n,q}$. From Proposition 1, we have

$$\begin{aligned}\phi_{n,q} &= \frac{1}{C_n}\sum_{j=0}^{n-1} f\left(\frac{2\pi j}{n}\right)e^{i\frac{2\pi qj}{n}} = \frac{1}{2\pi C_n}\sum_{p=-\infty}^{\infty}\phi_p\sum_{j=0}^{n-1}e^{i\frac{2\pi j}{n}(q-p)} \\ &= \frac{n}{2\pi C_n}\sum_{k=-\infty}^{\infty}\phi_{q-kn}.\end{aligned} \tag{2}$$

This characteristic function leads to the expressions for the qth cosine and sine moments as

$$\alpha_{n,q} = \frac{\alpha_q + \sum\limits_{k=1}^{\infty}\left(\alpha_{kn+q} + \alpha_{kn-q}\right)}{1 + 2\sum\limits_{k=1}^{\infty}\alpha_{nk}}, \quad \beta_{n,q} = \frac{\beta_q + \sum\limits_{k=1}^{\infty}\left(\beta_{kn+q} - \beta_{kn-q}\right)}{1 + 2\sum\limits_{k=1}^{\infty}\alpha_{nk}}.$$

3 Results

In this section, we show specifically how five families of DCDs are constructed from the corresponding CCDs. We use discrete von Mises, cardioid, nonnegative trigonometric sums [Fernández-Durán (2004)], wrapped Cauchy and four-parameter wrapped Cauchy [Kato and Jones (2015)] distributions to illustrate the construction. Similarly, other discrete circular distributions can be constructed.

3.1 Discrete von mises distribution

The pdf of a von Mises distribution with mean direction μ and concentration κ, VM(μ, κ), is

$$f_{\mathrm{VM}}(\theta) = \frac{1}{2\pi I_0(\kappa)}e^{\kappa\cos(\theta-\mu)},$$

for $\kappa \geq 0$ and $0 \leq \mu < 2\pi$. The pth cosine and sine moments are expressed by $\alpha_p = I_p(\kappa)\cos(p\mu)\,/I_0(\kappa)$ and $\beta_p = I_p(\kappa)\ \sin(p\mu)/I_0(\kappa)$ respectively, where $I_p(\cdot)$ is the modified Bessel function of the first kind and order p defined by

$$I_p(\kappa) = \frac{1}{2\pi}\int_0^{2\pi}\cos(p\theta)e^{\kappa\cos\theta}d\theta.$$

The pmf of the corresponding discrete von Mises distribution is expressed by

$$\begin{aligned}p_{\mathrm{VM}}(j) &= \frac{e^{\kappa\cos(2\pi j/n-\mu)}}{\sum_{p=0}^{n-1}e^{\kappa\cos(2\pi p/n-\mu)}}, \\ &= \frac{e^{\kappa\cos(2\pi j/n-\mu)}}{n\{I_0(\kappa) + 2\sum_{p=1}^{\infty}I_{np}(\kappa)\cos(np\mu)\}}, \quad j=0,1,2,\ldots,n-1,\end{aligned}$$

which involves Bessel functions in the normalizing constant. This distribution is denoted by DVM$_n(\mu, \kappa)$ or DVM$_n$ for short. For calculating this pmf numerically, the first expression is useful for small n, while the second one is useful for large n since $I_{np}(\kappa)<e^{0.25z^2}$ $(0.5z)^{np}/\Gamma(np+1)$ and it converges to 0 quickly as $p \to \infty$. From (2), the trigonometric moments are expressed as

$$\begin{aligned}\phi_{n,q} &= \frac{\sum_{j=0}^{n-1} e^{\kappa\cos(2\pi j/n-\mu)+i\frac{2\pi qj}{n}}}{\sum_{j=0}^{n-1} e^{\kappa\cos(2\pi j/n-\mu)}} \quad (\text{mod } q) \\ &= \frac{I_q(\kappa)e^{iq\mu} + \sum_{p=1}^{\infty}\{I_{pn-q}e^{-i(pn-q)\mu} + I_{pn+q}e^{i(pn+q)\mu}\}}{I_0(\kappa) + 2\sum_{p=1}^{\infty} I_{pn}(\kappa)\cos(pn\mu)} \quad (\text{mod } q).\end{aligned}$$

3.2 Discrete cardioid distribution

The pdf of a cardioid distribution with mean direction μ and mean resultant length ρ is

$$f_C(\theta) = \frac{1}{2\pi}\{1 + 2\rho\cos(\theta - \mu)\} = \frac{1}{2\pi}\{1 + 2\rho(\cos\mu\cos\theta + \sin\mu\sin\theta)\}$$

for $0 \le \mu < 2\pi$ and $0 \le \rho \le 1/2$. The pth cosine and sine moments are $\alpha_p = \rho\cos\mu$ and $\beta_p = \rho\sin\mu$ for $p=1$ and $\alpha_p = \beta_p = 0$ for $p = 2, 3, 4, \ldots$. The pmf of the corresponding discrete cardioid distribution is expressed by

$$p_C(j) = \frac{1}{n}\left\{1 + 2\rho\cos\left(\frac{2\pi j}{n} - \mu\right)\right\}, \quad j = 0, 1, 2, \ldots, n-1,$$

and is denoted by DC$_n(\mu, \rho)$ or DC$_n$ for short. The cumulative sum of $p_C(j)$ is also expressed by a closed form as

$$\sum_{j=0}^{k} p_C(j) = \frac{1}{n}\left[k + 1 + \frac{2\rho\sin\{\pi(k+1)/n\}\cos(\pi k/n - \mu)}{\sin(\pi/n)}\right], \quad k = 0, 1, 2, \ldots, n-1.$$

The trigonometric moments are

$$\phi_{n,q} = \begin{cases} 1, & q = 0, \\ \rho e^{\pm i\mu}, & q = \pm 1, \\ 0, & \text{otherwise}, \end{cases} \quad (\text{mod}\, n),$$

from which we observe that if $\Theta_{n,1} \sim$ DC$_n(\mu_1, \rho_1)$ and $\Theta_{n,2} \sim$ DC$_n(\mu_2, \rho_2)$ are independent, then $\Theta_{n,1} + \Theta_{n,2} \sim$ DC$_n(\mu_1 + \mu_2, \rho_1\rho_2)$.

3.3 Discrete nonnegative trigonometric sums distribution

A family of nonnegative trigonometric sums distributions with order M, denoted as NTS$_M$, is defined as

$$f_{\mathrm{NTS}}(\theta) = \frac{1}{2\pi}\left[1 + 2\sum_{m=1}^{M}\{a_m\cos(m\theta) + b_m\sin(m\theta)\}\right],$$

where $a_m = \sum_{v=0}^{M-m}\{r_{v+m}r_v + c_{v+m}c_v\}$ and $b_m = \sum_{v=0}^{M-m}\{r_{v+m}c_v - r_v c_{v+m}\}$ for $m = 1, 2, 3, \ldots, M$ with real numbers r_m and c_m such that $\sum_{m=0}^{M}(r_m^2 + c_m^2) = 1$. To estimate the parameters, some constraint, like $c_0 = 0$, is imposed in order to make the parameters of the model identifiable [Fernández-Durán (2007)].

Since the pth cosine and sine moments of NTS$_M$ are $\alpha_p = a_p$ and $\beta_p = b_p$ for $p = 1, 2, 3, \ldots, M$ and $\alpha_p = \beta_p = 0$ for $p > M$, the pmf of the corresponding discrete nonnegative trigonometric sums distribution with order M and division n is immediately obtained as

$$p_{\mathrm{NTS}}(j) = \frac{1}{n}\left[1 + 2\sum_{m=1}^{M}\left\{a_m\cos\left(\frac{2m\pi j}{n}\right) + b_m\sin\left(\frac{2m\pi j}{n}\right)\right\}\right], \quad j = 0, 1, 2, \ldots, n-1$$

and denoted by DNTS$_{M,n}(a_m, b_m; m = 1, 2, 3, \ldots, M)$ or DNTS$_{M,n}$ for short. DNTS$_{M,n}$ has $2M$-parameters a_m and b_m $(m = 1, 2, 3, \ldots, M)$, and hereafter we suppose that the number of the equidistant points on the circle (n) is greater than or equal to $2M$, or $M \leq n/2$, for model identifiability.

From Proposition 1, the cumulative sum of $p_{\mathrm{NTS}}(j)$ is expressed by

$$\sum_{j=0}^{k} p_{NTS}(j) = \frac{1}{n}\left[k + 1 + \frac{2\sin\{\pi(k+1)/n\}}{\sin(\pi/n)} \times \sum_{m=1}^{M}\left\{a_m\cos\left(\frac{m\pi k}{n}\right) + b_m\sin\left(\frac{m\pi k}{n}\right)\right\}\right], \quad k = 0, 1, 2, \ldots, n-1.$$

The trigonometric moments are

$$\phi_{n,q} = \begin{cases} a_q + ib_q, & 1 \leq q \leq M, \\ 0, & M < q < n - M, \\ a_{n-q} - ib_{n-q}, & n - M \leq q \leq n - 1. \end{cases}$$

Note that the trigonometric moments of DNTS$_{M,n}$ are the same as those of NTS$_M$. This means that the discretization from NTS$_M$ to DNTS$_{M,n}$ preserves the mean direction and the mean resultant length. When $n = 2M$, the moments are expressible by

$$\phi_{n,q} = \begin{cases} a_q + ib_q, & 1 \le q < M, \\ a_M, & q = M, \\ a_{2M-q} - ib_{2M-q}, & M < q \le 2M-1. \end{cases}$$

This family includes discrete cardioid distribution (DC$_n$) as a special case.

3.4 Discrete wrapped cauchy distribution

The pdf of a wrapped Cauchy distribution with mean direction μ and mean resultant length ρ is given by

$$f_{\mathrm{WC}}(\theta) = \frac{1-\rho^2}{2\pi\{1+\rho^2-2\rho\cos(\theta-\mu)\}}$$

for $0 \le \mu < 2\pi$ and $0 \le \rho < 1$, and the pth cosine and sine moments are $\alpha_p = \rho^p \cos(p\mu)$ and $\beta_p = \rho^p \sin(p\mu)$ for $p = 1, 2, 3, \ldots$.

Since

$$\sum_{p=1}^{\infty} \alpha_{np} = \mathrm{Re}\left[\sum_{p=1}^{\infty} \rho^{np} e^{inp\mu}\right] = \frac{\rho^n\{\cos(n\mu) - \rho^n\}}{1+\rho^{2n}-2\rho^n\cos(n\mu)},$$

the pmf of the corresponding discrete wrapped Cauchy distribution, denoted by DWC(μ, ρ) or DWC$_n$ for short, is expressed by

$$p_{\mathrm{WC}}(j) = \frac{(1-\rho^2)\{1+\rho^{2n}-2\rho^n\cos(n\mu)\}}{n(1-\rho^{2n})\{1+\rho^2-2\rho\cos(2\pi j/n-\mu)\}}, \quad j = 0, 1, 2, \ldots, n-1$$

and its trigonometric moments are

$$\phi_q = \frac{1}{1-\rho^{2n}}\left\{(\rho^{n-q} - \rho^{n+q})e^{i(q-n)\mu} + (\rho^q - \rho^{2n-q})e^{iq\mu}\right\} \quad (\mathrm{mod}\ n).$$

3.5 Discrete four-parameter wrapped cauchy distribution

The pdf of a four-parameter wrapped Cauchy (FWC) distribution is

$$f_{\mathrm{FWC}}(\theta) = \frac{1}{2\pi}\left\{1 + 2\gamma\frac{\cos(\theta-\mu) - \rho\cos\lambda}{1+\rho^2-2\rho\cos(\theta-\mu-\lambda)}\right\}$$

for $0 \le \mu < 2\pi$, $0 \le \gamma < 1$, $0 \le \rho < 1$, and $-\pi \le \lambda < \pi$, where $(\rho\cos\lambda - \gamma)^2 + (\rho\sin\lambda)^2 \le (1-\gamma)^2$. In this model, μ and γ play the roles of location and concentration respectively, and λ and ρ controls skewness and kurtosis of the distribution [Kato and Jones (2015)]. As a special case, this distribution reduces to the wrapped Cauchy distribution when λ=0 and γ=ρ. The pth cosine and sine moments are $\alpha_p = \gamma\rho^{p-1}\cos\{p(\mu+\lambda)-\lambda\}$ and $\beta_p = \gamma\rho^{p-1}\sin\{p(\mu+\lambda)-\lambda\}$, respectively.

Since

$$\sum_{p=1}^{\infty} \gamma\rho^{np-1} e^{i\{np(\mu+\lambda)-\lambda\}} = \frac{\gamma\rho^{n-1}\{e^{in(\mu+\lambda)-\lambda} - \rho^n e^{i\lambda}\}}{1+\rho^{2n} - 2\rho^n\cos\{n(\mu+\lambda)\}},$$

taking its real part, we have

$$\sum_{p=1}^{\infty} \alpha_{np} = \frac{\gamma\rho^{n-1}[\cos\{n(\mu+\lambda)-\lambda\} - \rho^n\cos\lambda]}{1+\rho^{2n} - 2\rho^n\cos\{n(\mu+\lambda)\}}.$$

Thus, the pmf of the corresponding four-parameter discrete wrapped Cauchy distribution is expressed by

$$\begin{aligned} &p_{\mathrm{FWC}}(j) \\ &\quad = 1+\rho^{2n} - 2\rho^n\cos\{n(\mu+\lambda)\} \\ &\qquad \div\, n\big(1+\rho^{2n} - 2\gamma\rho^{2n-1}\cos\lambda + 2\rho^{n-1}[\gamma\cos\{n(\mu+\lambda)-\lambda\} - \rho\cos\{n(\mu+\lambda)\}]\big) \\ &\qquad \times \left\{1 + 2\gamma\frac{\cos(2\pi j/n-\mu) - \rho\cos\lambda}{1+\rho^2 - 2\rho\cos(2\pi j/n-\mu-\lambda)}\right\}, \quad j = 0,1,2,\ldots,n-1 \end{aligned}$$

and is denoted by $\mathrm{DFWC}_n(\mu,\rho,\lambda,\gamma)$ or DFWC_n for short. It can be shown that plugging λ=0 and γ=ρ into $p_{\mathrm{FWC}}(j)$ results in $p_{\mathrm{WC}}(j)$ of the discrete wrapped Cauchy distribution. The trigonometric moments are omitted to save space.

4 Estimation and simulation

In this section, we utilize two methods to estimate the parameters of the constructed DCDs. Next, we compare these methods under different sample sizes and number of divisions via a simulation study.

4.1 Estimation

Maximum likelihood based inference for grouped circular data is mentioned in Pewsey et al. [Pewsey, Neuhäuser and Ruxton (2013), Section 6.3.5]. As a special feature of our discretization method, however, there is a relationship between the log-likelihood functions for the base CCD and the corresponding DCD as follows.

Let f_j, j=0, 1, 2,…, n−1, be the frequency observed at the value $2\pi j/n$ on a circumference and N=$\sum_{j=0}^{n-1} f_j$ be the sample size of the dataset. Then, the log-likelihood function of DCD in (1) with parameter vector ψ for the dataset is given by

$$L(\psi) = \sum_{j=0}^{n-1} f_j \log p_n(j;\psi) = \sum_{j=0}^{n-1} f_j \log f(2\pi j/n;\psi) - N\log C_n(\psi),$$

which is equal to the log-likelihood function of the base CCD subtracting the penalty term $N\log C_n(\psi)$ for the DCD. From this expression, the maximum likelihood estimation (MLE) of DCD does not correspond to that of its base continuous circular distribution in general. An alternative approach to estimate parameters of DCD is the minimum chi-square estimation (MCSE), where the estimates are given by minimizing Pearson's chi-square test statistic

$$CS(\psi) = \sum_{j=0}^{n-1} \frac{\{f_j - Np_n(j)\}^2}{Np_n(j)}.$$

4.2 Simulation

We conducted a simulation study to compare how well the original CCDs and the corresponding DCDs model grouped data, in which various sample sizes N were generated with 5,000 Monte Carlo repetitions.

For a continuous distribution whose distribution function is not analytic, it is difficult to generate random numbers directly; see Appendix A for further explanation. Thus, we first generated random numbers from DCD with large divisions m=1,152 (=9×2^7) using the method in Appendix A, then we grouped them into equidistant division n=9, 18 and 36 of [0, 2π), each number in [2$\pi j/n$, 2π (j+1)/n) was rounded to (2j+1)π/n. The number of divisions 36 and 18 were chosen, because the number of orthologs in every 10° and 20° arcs were of interest in the Application section, while n=9 was studied as a small number of divisions.

From these rounded numbers, we estimated the parameters of the base CCD and the corresponding DCD by MLE and MCSE in the Estimation section and compared the fitting performance.

As measures of fitting performance, we use the chi-square test statistic

$$\chi^2 = \sum_{j=0}^{n-1} \frac{(f_j - np_j)^2}{np_j}$$

and Freedman's test statistic [Freedman (1981)], which is a modified version of Watson's U^2 statistic for discrete data,

$$U^{*2} = \frac{N}{n} \sum_{j=0}^{n-1} \left\{ \sum_{j=0}^{n-1} S_j^2 - \left(\sum_{j=0}^{n-1} S_j \right)^2 \right\}$$

where

$$S_j = \sum_{k=0}^{j} \left(\frac{f_k}{N} - p_k \right), \quad j = 0, 1, 2, \ldots, n-1$$

and

$$p_j = \begin{cases} p_n(j) & \text{for DCD} \\ \displaystyle\int_{2\pi j/n}^{2\pi(j+1)/n} f(\theta)d\theta & \text{for CCD}, \quad j=0,1,2,\ldots,n-1. \end{cases}$$

Because the exact distributions of χ^2 and U^{*2} are intractable, the exact p-values are difficult to calculate. Thus, we calculated the approximated p-values by the bootstrap method. Specifically, we generated bootstrap samples $\chi^{2(b)}$ of χ^2 and $U^{*2(b)}$ of U^{*2}, b=1, 2, 3, …, B and B=10,000, under the true parameters. Next, we calculated bootstrapped p-values $p^B_{\chi^2} := \#\{\chi^{2(b)} > x$, b=1, 2, 3, …, $B\}/B$ and $p^B_{U^{*2}} := \#\ \{U^{\ *2(b)} > u$, b=1, 2, 3, …, $B\}/B$, where $\#A$ is the number of elements in the set A, and x and u are the realizations of the chi-square and Freedman's test statistics obtained from the estimated distribution, respectively.

We conducted the simulation for discrete cardioid, wrapped Cauchy and four-parameter wrapped Cauchy distributions. The results are very similar, so we only show the result of discrete four-parameter wrapped Cauchy distributions and omit the others. Tab. 1 shows the bias and mean squared error (MSE) of the estimates by MLE for FWC$_n$(π, 0.6, π/2, 0.3) and those by MLE and MCSE for DFWC$_n$(π, 0.6, π/2, 0.3) with their measures of fitting performance $p^B_{\chi^2}$ and $p^B_{U^{*2}}$.

The original CCD fits the data as well as the corresponding DCD, only when the sample size is small (N=50) and the number of divisions is large (n=36). While in other cases, the DCD performs better than its corresponding CCD in terms of $p^B_{\chi^2}$ and $p^B_{U^{*2}}$. This result is natural since large n leads to a continuous sample while small n leads to a discrete sample.

Comparing the estimations, we see that when the number of divisions is large (n=36 for N ranging from 50 to 200), MLE is better than MCSE in the sense of $p^B_{U^{*2}}$, small bias and MSE of the estimated parameters. However, when the number of divisions and sample size are not large (n=9 and 18; N=50 and 100, respectively), MCSE outperforms MLE in most cases for estimating parameters and fitting to discrete circular data.

Our algorithm was written in *Mathematica* 8.0 and we optimized the functions $L(\psi)$ and CS (ψ) in Section 4.1 by the *NMaximize* and *NMinimize* commands of the Nelder-Mead method. The running time (CPU: Core i7-5820K, 3.30 GHz) of the simulation was very fast. For N=50, 100 and 200 simulated data with the number of divisions n=9, 18 and 36, the running time was bounded by data grouped by 36 divisions. Of these, applying MLE and MCSE to DFWC$_{36}$ took 2.0 and 0.1 seconds, respectively, while applying MLE to FWC took about 2.8 seconds.

5 Application

Most prokaryotic (bacteria and archaea) genomes are made up of single circular chromosomes, called circular genomes. Orthologs are genes in different species that evolved directly from a common ancestral gene by speciation, and they can be used to

Table 1: Comparisons of estimates for the parameters in the original and discretized FWC (π, 0.6, π/2, 0.3)

N			**36 divisions**			**18 divisions**			**9 divisions**		
			MLE FWC	**MLE DFWC$_{36}$**	**MCSE DFWC$_{36}$**	**MLE FWC**	**MLE DFWC$_{18}$**	**MCSE DFWC$_{18}$**	**MLE FWC**	**MLE DFWC$_9$**	**MCSE DFWC$_9$**
50	μ	Bias	0.037	0.035	−0.059	0.041	0.028	−0.031	0.103	0.066	0.025
		MSE	0.111	0.112	0.129	0.117	0.115	0.128	0.138	0.200	0.198
	ρ	Bias	0.005	0.007	0.009	−0.002	−0.003	−0.004	0.028	−0.015	−0.019
		MSE	0.029	0.029	0.030	0.031	0.028	0.030	0.050	0.035	0.036
	λ	Bias	−0.135	−0.130	−0.035	−0.144	−0.125	−0.087	−0.217	−0.113	−0.079
		MSE	0.457	0.457	0.659	0.501	0.510	0.646	0.544	0.655	0.712
	γ	Bias	0.017	0.018	−0.026	0.013	0.023	−0.002	−0.014	0.040	0.030
		MSE	0.008	0.007	0.006	0.008	0.008	0.006	0.010	0.014	0.012
	$p^B_{\chi^2}$		0.671	0.672	0.715	0.691	0.711	0.735	0.629	0.787	0.799
	$p^B_{U^{*2}}$		0.824	0.827	0.734	0.799	0.840	0.815	0.675	0.861	0.860
100	μ	Bias	0.012	0.011	−0.041	0.017	0.008	−0.028	0.140	0.018	−0.007
		MSE	0.049	0.048	0.055	0.057	0.052	0.053	0.093	0.082	0.082
	ρ	Bias	0.000	0.001	0.002	−0.007	−0.010	−0.010	0.094	−0.019	−0.021
		MSE	0.014	0.014	0.015	0.017	0.015	0.016	0.050	0.020	0.020
	λ	Bias	−0.036	−0.033	0.025	−0.039	−0.029	0.008	−0.226	−0.026	0.011
		MSE	0.156	0.153	0.220	0.204	0.199	0.218	0.302	0.316	0.328
	γ	Bias	0.009	0.009	−0.020	0.005	0.009	−0.006	−0.048	0.014	0.008
		MSE	0.004	0.003	0.003	0.004	0.004	0.003	0.011	0.005	0.004
	$p^B_{\chi^2}$		0.649	0.656	0.688	0.683	0.707	0.723	0.483	0.793	0.801
	$p^B_{U^{*2}}$		0.817	0.817	0.751	0.803	0.828	0.816	0.487	0.867	0.867
200	μ	Bias	0.003	0.002	−0.023	0.006	0.003	−0.016	0.178	−0.002	−0.018
		MSE	0.024	0.024	0.025	0.027	0.025	0.025	0.072	0.037	0.036
	ρ	Bias	−0.002	−0.002	0.000	0.010	−0.012	−0.011	0.148	−0.021	−0.022
		MSE	0.006	0.006	0.006	0.007	0.006	0.007	0.053	0.011	0.011
	λ	Bias	−0.004	−0.002	0.029	−0.005	−0.002	0.016	−0.254	0.009	0.036
		MSE	0.067	0.066	0.074	0.074	0.075	0.079	0.153	0.152	0.151
	γ	Bias	0.004	0.004	−0.013	0.002	0.003	−0.004	−0.074	0.003	0.000
		MSE	0.002	0.002	0.002	0.002	0.002	0.002	0.011	0.002	0.002
	$p^B_{\chi^2}$		0.653	0.646	0.669	0.685	0.696	0.706	0.290	0.794	0.799
	$p^B_{U^{*2}}$		0.803	0.809	0.767	0.799	0.811	0.807	0.286	0.866	0.866

compare structures of paired bacterial genomes [Shieh, Zheng, Johnson et al. (2011)]. Comparisons of structures, e.g., the most and least favoured spots, between paired genomes are important and useful in the area of synthetic and evolutionary biology. The data were downloaded from NCBI (ftp://ftp.ncbi.nlm.gov/refseq/release/bacteria), and were preprocessed as stated in Shieh et al. [Shieh, Zheng, Johnson et al. (2011)] The processed data are available at http://www.stat.sinica.edu.tw/gshieh/DCDs/data.txt.

When comparing the genomes of paired bacteria, e.g., *Clostridium* and *Sulfolobus*, whose genomes were plotted at http://www.stat.sinica.edu.tw/gshieh/DCDs/genomes.pdf, in the different kingdoms, their shared orthologs are discrete when depicted in binned circumferences on unit circles. Fig. 1 shows rose diagrams of the shared orthologs between *Clostridium* and *Sulfolobus*. The distributions of the shared orthologs in both genomes look different, because they have evolved to belong to different kingdoms. Henceforth, we call the binned shared orthologs in *Clostridium* and *Sulfolobus* the *Clostridium* and *Sulfolobus* data, respectively, which are sometimes presented in degrees on the circumference of a unit circle, for clarity of depiction. The total number of shared orthologs N is 192 for both data sets. We use 18 pre-fixed equidistant (20°) bins and center the number of shared orthologs in each bin between $20j$ and $20(j+1)$ degrees for $j=0,1,2,\dots,17$, namely at $10°, 30°, 50°, \dots, 330°$. For the *Clostridium* data, the sample mean direction is $\bar{\theta}=0.52$ (in radian) and the sample mean resultant length is $\bar{R}=0.19$ with p-value 0.001 of the Rayleigh test for uniformity based on $2N\bar{R}^2=13.41$ which means that the *Clostridium* data is not uniform on the circle; for the *Sulfolobus* data, they are $\bar{\theta}=0.18$ and $\bar{R}=0.14$ with $2N\bar{R}^2=7.03$ and p-value 0.03, respectively. Fig. 1 shows

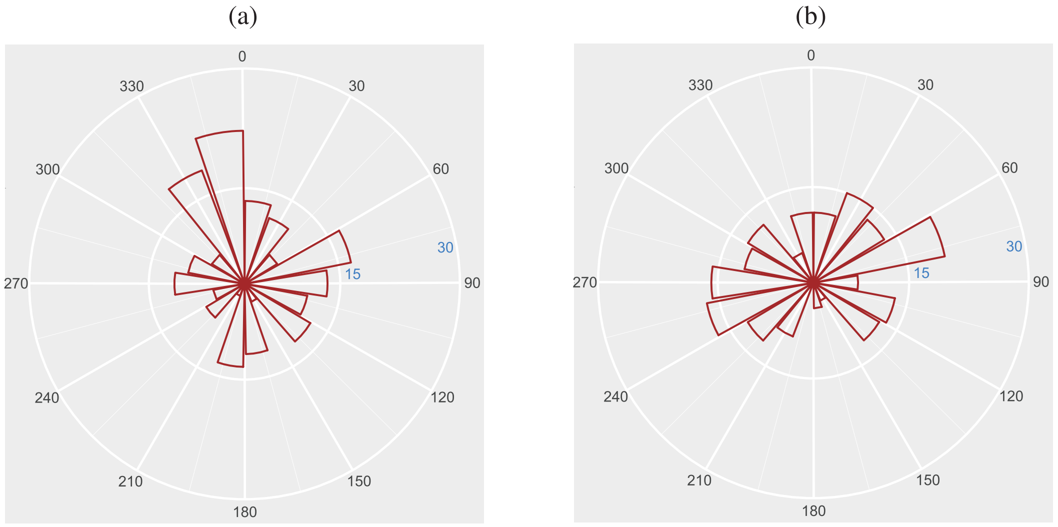

Figure 1: Rose diagrams of (a) *Clostridium* and (b) *Sulfolobus* data with 18 pre-fixed equidistant (20°) points

that the distribution of the *Clostridium* data looks asymmetric and has the mode around 350°, followed by 330° and 70°, while the *Sulfolobus* data looks symmetric and is multimodal with modes around 70°, 250° and 270°, respectively.

When fitting distributions to the *Clostridium* and *Sulfolobus* data, we apply the MCSE to estimate the parameters of DVM_n, DC_n, $\mathrm{DNTS}_{2,n}$, DWC_n and DFWC_n. In addition, we also fit uniform distribution and two recently studied DCDs (IWP and WDSL), and compare their performances to those of our constrcuted DCDs. IWP assumes the following pmf.

$$P_{\mathrm{IWP}}(j) = \sum_{k=0}^{\infty} \frac{\lambda^{\delta(j-n\mu/(2\pi))+kn}}{(\delta(j-\mu/(2\pi))+kn)!} e^{-\lambda}, \quad j = 0, 1, 2, \ldots, n-1,$$

where $\lambda>0$, $\delta \in \{-1, 1\}$, $\mu \in \left\{0, \frac{2\pi}{n}, \frac{4\pi}{n}, \frac{6\pi}{n}, \ldots, \frac{2\pi(n-1)}{n}\right\}$. To apply this distribution, we need to approximate the infinite summation by truncation, and the truncation point depends on the parameter λ. For stable calculation, we restrict the parameter space of λ to $0<\lambda<60$ and fix the truncation point at 100.

The WDSL distribution (Jayakumar and Jacob, 2012) has the following pmf.

$$P_{\mathrm{WDSL}}(j) = \frac{(1-p)(1-q)}{1-pq}\left[\frac{q^{n-j}(1-p^n)+p^j(1-q^n)}{(1-p^n)(1-q^n)}\right], \quad j = 0, 1, 2, \ldots, n-1,$$

where $0<p<1$, $0<q<1$. Since this distribution does not have the location parameter and the periodicity, i.e., $P_{\mathrm{WDSL}}(j) \neq P_{\mathrm{WDSL}}(j+n)$, we use the generalized WDSL (denoted by gWDSL) which has the following pmf.

$$P_{\mathrm{gWDSL}}(j) = \begin{cases} P_{\mathrm{WDSL}}\left(j - \frac{n\mu}{2\pi} + n\right), & j < \frac{n\mu}{2\pi}, \\ P_{\mathrm{WDSL}}\left(j - \frac{n\mu}{2\pi}\right), & j \geq \frac{n\mu}{2\pi}, \end{cases}$$

where $\mu \in \left\{0, \frac{2\pi}{n}, \frac{4\pi}{n}, \frac{6\pi}{n}, \ldots, \frac{2\pi(n-1)}{n}\right\}$.

Tabs. 2 and 3 show the MCSE for the parameters of each distribution, model selection criterion AIC and bootstrapped p-value of the Freedman's goodness-of-fit statistic $p^B_{U^{*2}}$ for the *Clostridium* and *Sulfolobus* data, respectively. In the sense of AIC and $p^B_{U^{*2}}$, DFWC_{18} gives the best fit for the *Clostridium* data with the smallest AIC and p-value 0.629. For the *Sulfolobus* data $\mathrm{DNTS}_{2,18}$ gives the best fit with the smallest AIC and the largest p-value 0.882 (the significance level equal to 0.05). Figs. 2 and 3 show the linear histograms of the *Clostridium* data and *Sulfolobus* data, respectively, superposed by well-fitted distributions as line plots. As shown in Fig. 3, $\mathrm{DNTS}_{2,18}$ models the *Sulfolobus* data better than the remaining ones.

Table 2: Fitting DCDs to the *Clostridium* data by MCSE at 18 pre-fixed equidistant points

Distribution	**MCSE**	**AIC**	$\boldsymbol{p^{B}_{U^{*2}}}$
Uniform		1109.90	0.0004
DC_{18}	$\hat{\mu}=0.4290$, $\hat{\rho}=0.1699$	1100.36	0.3666
DVM_{18}	$\hat{\mu}=0.3244$, $\hat{\rho}=0.3554$	1100.37	0.3958
DWC_{18}	$\hat{\mu}=0.2176$, $\hat{\rho}=0.1760$	1100.52	0.3762
DFWC_{18}	$\hat{\mu}=0.3109$, $\hat{\rho}=0.8340$ $\hat{\lambda}=5.5112$, $\hat{\gamma}=0.1497$	1099.11	0.6290
$\mathrm{DNTS}_{2,18}$	$\hat{a}_1=0.1603$, $\hat{b}_1=0.0674$ $\hat{a}_2=0.0745$, $\hat{b}_2=-0.0611$	1102.94	0.6394
IWP	$\hat{\mu}=2.7925$, $\hat{\lambda}=29.3143$ $\hat{\delta}=1$	1131.57	0.0000
gWDSL	$\hat{\mu}=5.5851$, $\hat{\rho}=0.9418$ $\hat{q}=0.0100$	1163.06	0.0000

Table 3: Fitting DCDs to the *Sulfolobus* data by MCSE at 18 pre-fixed equidistant points

Distribution	**MCSE**	**AIC**	$\boldsymbol{p^{B}_{U^{*2}}}$
Uniform		1109.90	0.0034
DC_{18}	$\hat{\mu}=3.7723$, $\hat{\rho}=0.0000$	1113.90	0.0032
DVM_{18}	$\hat{\mu}=0.0260$, $\hat{\rho}=0.1991$	1106.84	0.0636
DWC_{18}	$\hat{\mu}=0.1606$, $\hat{\rho}=0.0834$	1107.96	0.0428
DFWC_{18}	$\hat{\mu}=6.0284$, $\hat{\rho}=0.6698$ $\hat{\lambda}=3.6262$, $\hat{\gamma}=0.1731$	1096.10	0.3008
$\mathrm{DNTS}_{2,18}$	$\hat{a}_1=0.1310$, $\hat{b}_1=0.0294$ $\hat{a}_2=-0.2085$, $\hat{b}_2=0.1014$	1086.66	0.8824
IWP	$\hat{\mu}=0.6981$, $\hat{\lambda}=33.9697$ $\hat{\delta}=1$	1109.41	0.0348
gWDSL	$\hat{\mu}=1.0472$, $\hat{\rho}=0.4768$ $\hat{q}=0.9532$	1115.87	0.0212

Furthermore, in terms of running time (CPU: Core i7-5820K, 3.30 GHz), our constructed DCDs took less than 0.5 seconds, while IWP and generalized WDSL took 900 and five seconds, respectively. This is because IWP and generalized WDSL have discrete location

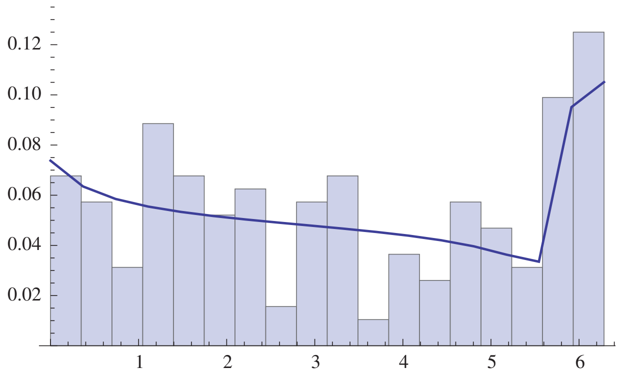

Figure 2: Linear histogram of the *Clostridium* Data superposed by the fitted DFWC$_{18}$ (line plot) at 18 pre-fixed equidistant points

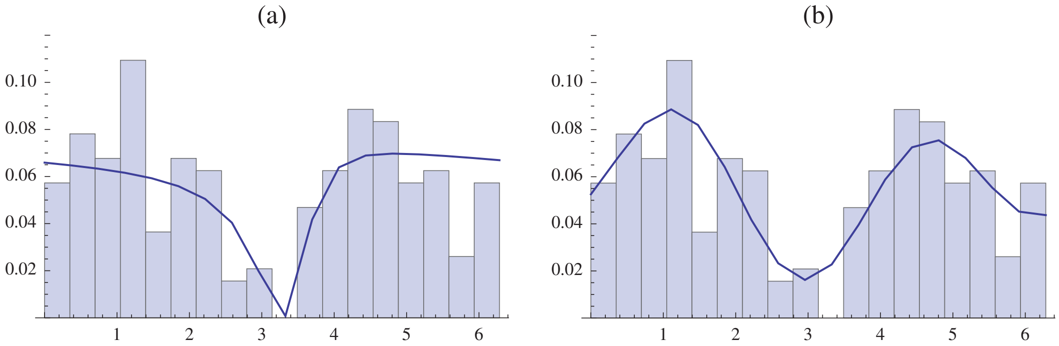

Figure 3: Linear histograms of the *Sulfolobus* Data superposed by the fitted (a) DFWC$_{18}$ and (b) DNTS$_{2,\ 18}$ at 18 pre-fixed equidistant points

parameters, which are estimated by profile MLE like method. In addition, IWP has a sum of many terms.

We summarize the modeling of the *Clostridium* and *Sulfolobus* data by DCDs as follows.

1. For the *Clostridium* data, asymmetric DCDs yield good fits and, among these, the DFWC$_{18}$ yields the best fit in terms of AIC and Freedman's goodness-of-fit test.
2. For the *Sulfolobus* data, DNTS$_{2,18}$ gives the best fit in terms of AIC and Freedman's goodness-of-fit test. Fig. 3 also shows that DNTS$_{2,18}$ fits the data better than the second best DFWC$_{18}$, in terms of the line plots fitting to the linear histogram of the data. Note that DFWC$_{18}$ can not exhibit bimodality.
3. The above results suggest that these shared orthologs are distributed differently, which is consistent with the fact that these prokaryotes belong to different kingdoms. Nevertheless, the rose diagrams (Fig. 1) show that these shared orthologs have a

common favored region (near 70°) out of two to three favored regions. This finding is important for building synthetic prokaryotic genomes in synthetic biology.

6 Conclusions

We have investigated the construction of DCDs by generating the pmfs from circular pdfs. The normalizing constant is simply represented by the cosine moments of the base CCD, and the constructed discrete distributions are tractable. Simulation studies show that DCDs outperform the corresponding CCDs in modeling grouped (discrete) circular data, and MCSE is better than MLE. The constructed DCDs, IWP and WDSL were applied to compare the structures of shared orthologs in a pair of prokaryotes, an important topic in synthetic and evolutionary biology. Specifically, DFWC (DNTS) distribution is shown to model multi-modal shared orthologous genes in bacteria *Clostridium* (archaea *Sulfolobus*) well. We conclude that of the distributions considered, DFWC_n fits the asymmetric *Clostridium* data the best in terms of AIC and Freedman's goodness-of-fit test, and $\mathrm{DNTS}_{2,n}$ fits the symmetric and multi-modal *Sulfolobus* data better than the remaining DCDs considered. Although these shared orthologs followed different DCDs, they do have a common favored region around 70° out of two to three top-favored regions. This finding is important for building synthetic prokaryotic genomes in synthetic biology.

The constructed DCDs are versatile, namely these families of distributions can fit both uni-modal and multi-modal and symmetric and asymmetric discrete circular data. Moreover, the computation of our algorithm, consisting of estimation of the parameters and goodness of fit test, is very fast. The presented discretization method can be applied to any univariate CCDs, but it is limited to univariate circular distributions. Nevertheles, our method is the basis for bivariate DCDs which is of interest and have applications in many scientific areas. Therefore, we leave construction of discrete bivariate circular models for future study.

Acknowledgement: We thank three anonymous reviewers for constuctive comments and suggestions which helped improve the article.

Funding Statement: Tomoaki Imoto was supported by JSPS KAKENHI Grant Number 18K13459 and Grace S. Shieh was supported in part by MOST 106-2118-M-001-017 and MOST 107-2118-M-001-009-MY2.

Conflicts of Interest: The authors declare that there is no conflicts of interest regarding the present study.

References

Carrera, J.; Rodrigo, G.; Jaramillo, A. (2009): Model-based redesign of global transcription regulation. *Nucleic Acids Research*, vol. 37, no. 5, pp. e38. DOI 10.1093/nar/gkp022.

Fernández-Durán, J. (2004): Circular distributions based on nonnegative trigonometric sums. *Biometrika*, vol. 60, pp. 499-503.

Fernández-Durán, J. (2007): Models for circular-linear and circular-circular data constructed from circular distributions based on nonnegative trigonometric sums. *Biometrika*, vol. 63, pp. 579-585.

Fisher, N. (1993): *Statistical Analysis of Circular Data*. Cambridge University Press, Cambridge, UK.

Freedman, L. (1981): Watson's U_N^2 statistic for a discrete distribution. *Biometrika*, vol. 68, pp. 708-711.

Jacob, S.; Jayakumar, K. (2013): Wrapped geometric distribution: a new probability model for circular data. *Journal of Statistical Theory and Applications*, vol. 12, no. 4, pp. 348-355. DOI 10.2991/jsta.2013.12.4.3.

Jammalamadaka, S.; SenGupta, A. (2001): *Topics in Circular Statistics*. World Scientific, Farrer Road, Singapore.

Jayakumar, K.; Jacob, S. (2012): Wrapped skew laplace distribution on integers: a new probability model for circular data. *Open Journal of Statistics*, vol. 2, no. 01, pp. 106-114. DOI 10.4236/ojs.2012.21011.

Johnson, R.; Wehrly, T. (1977): Measures and models for angular correlation and angular-linear correlation. *Journal of the Royal Statistical Society: Series B*, vol. 39, pp. 222-229.

Kato, S.; Jones, M. (2015): A tractable and interpretable four-parameter family of unimodal distributions on the circle. *Biometrika*, vol. 102, no. 1, pp. 181-190. DOI 10.1093/biomet/asu059.

Mardia, K.; Jupp, P. (2000): *Directional Statistics*. Wiley, Chichester, USA.

Mastrantonio, G.; Jona Lasinio, G.; Maruotti, A.; Calise, G. (2015). On initial direction, orientation and discreteness in the analysis of circular variables.

Pewsey, A.; Neuhäuser, M.; Ruxton, G. (2013): *Circular Statistics in R*. Oxford University Press, Oxford, UK.

Shieh, G.; Zheng, S.; Johnson, R.; Chang, Y.; Shimizu, K. et al. (2011): Modeling and comparing the organization of circular genomes. *Bioinformatics*, vol. 27, no. 7, pp. 912-918. DOI 10.1093/bioinformatics/btr049.

Tatusov, R.; Koonin, E.; Lipman, D. (1997): A genomic perspective on protein families. *Science*, vol. 278, pp. 631-637.

Appendix A. Random number generation for DCDs

The distribution function $F_m(k)$ of a DCD defined on m pre-fixed equidistant points is a step function with m steps. Therefore, we can easily use the inverse transform method to generate random number as $\frac{2\pi}{m}\min\{k; F_m(k) \geq y\}$, where y is the random number of continuous uniform distribution defined on (0, 1). When m is a large integer, the generated random number approximates that of the base continuous circular distribution. This fact is confirmed as follows.

Let $f(\theta)$ be the pdf of a base CCD and $p_m(j)$ be the pmf of the corresponding DCD defined on m pre-fixed equidistant points. Assuming that $f \in C^2(0, 2\pi)$, then we have

$$\sum_{k=0}^{[(m-1)a]} p_m(k) = \frac{\frac{2\pi}{m}\sum_{j=0}^{[(m-1)a]} f\left(\frac{2\pi j}{m}\right)}{\frac{2\pi}{m}\sum_{j=0}^{m-1} f\left(\frac{2\pi j}{m}\right)} = \int_0^{2\pi a} f(\theta)d\theta + O\left(\frac{1}{m^2}\right),$$

for any $0<a<1$, where $[x]$ denotes the greatest integer less than or equal to x since

$$\frac{2\pi}{m}\sum_{j=0}^{m-1} f\left(\frac{2\pi j}{m}\right) = \int_0^{2\pi} f(\theta)d\theta + O\left(\frac{1}{m^2}\right) = 1 + O\left(\frac{1}{m^2}\right)$$

and

$$\begin{aligned}
&\frac{2\pi}{m}\sum_{j=0}^{[(m-1)a]} f\left(\frac{2\pi j}{m}\right) \\
&= \frac{[(m-1)a]+1}{ma}\frac{2\pi a}{[(m-1)a]+1}\sum_{j=0}^{[(m-1)a]}\left\{f\left(\frac{2\pi aj}{[(m-1)a]+1}\right) + O\left(\frac{1}{m^2}\right)\right\} \\
&= \left\{1 + O\left(\frac{1}{m^2}\right)\right\}\left\{\int_0^{2\pi a} f(\theta)d\theta + O\left(\frac{1}{m^2}\right)\right\} \\
&= \int_0^{2\pi a} f(\theta)d\theta + O\left(\frac{1}{m^2}\right).
\end{aligned}$$

For a continuous distribution whose distribution function is not expressed by an elementary function, it is difficult to apply the inverse transform method. In such a case, the acceptance-rejection method is often applied. However, the acceptance-rejection method will take time to generate the random numbers in general and will reject many random numbers from the proposal distribution. In contrast, our method can be executed easily by finding the point where the distribution function (step function) is larger than the random variable of uniform distribution.