The Annals of Applied Statistics
2021, Vol. 15, No. 4, 1583–1603
https://doi.org/10.1214/21-AOAS1459
© Institute of Mathematical Statistics, 2021

CLUSTERING ON THE TORUS BY CONFORMAL PREDICTION

BY SUNGKYU JUNG[1,*], KIHO PARK[1,†] AND BYUNGWON KIM[2]

[1]*Department of Statistics, Seoul National University,* [*]*sungkyu@snu.ac.kr;* [†]*pkh503201@snu.ac.kr*
[2]*Department of Statistics, Kyungpook National University, byungwonkim@knu.ac.kr*

Motivated by the analysis of torsion (dihedral) angles in the backbone of proteins, we investigate clustering of bivariate angular data on the torus $[-\pi, \pi) \times [-\pi, \pi)$. We show that naive adaptations of clustering methods, designed for vector-valued data, to the torus are not satisfactory and propose a novel clustering approach based on the conformal prediction framework. We construct several prediction sets for toroidal data with guaranteed finite-sample validity, based on a kernel density estimate and bivariate von Mises mixture models. From a prediction set built from a Gaussian approximation of the bivariate von Mises mixture, we propose a data-driven choice for the number of clusters and present algorithms for an automated cluster identification and cluster membership assignment. The proposed prediction sets and clustering approaches are applied to the torsion angles extracted from three strains of coronavirus spike glycoproteins (including SARS-CoV-2, contagious in humans). The analysis reveals a potential difference in the clusters of the SARS-CoV-2 torsion angles, compared to the clusters found in torsion angles from two different strains of coronavirus, contagious in animals.

1. Introduction.

The structure of a protein is often summarized by the torsion (dihedral) angles formed along the backbone of the protein (Dill and MacCallum (2012)). A standard visual representation of the torsion angles is given by the Ramachandran plot in which the sequence of (ϕ, ψ) torsion angles extracted from the backbone of a protein is plotted on $[-\pi, \pi) \times [-\pi, \pi)$ (Figure 1). Due to the unique (circular) challenges in the relatively simple data structure and to its importance in structural analysis of proteins and RNA, there have been a number of endeavors on density estimation, clustering and dimension reduction of data on the torus (Mardia, Taylor and Subramaniam (2007), Mardia et al. (2008, 2012), Eltzner, Huckemann and Mardia (2018), Gao et al. (2018), Lennox et al. (2009), Shapovalov, Vucetic and Dunbrack Jr. (2019), Nodehi et al. (2021)). In particular, clusters in the torsion angles have been interpreted as local structures of the backbone of a protein which determine the protein's functions (Berg, Tymoczko and Stryer (2002)).

The circular nature of angles leads that the bivariate angle (ϕ, ψ) is on the torus, which may be embedded as an intrinsically two-dimensional manifold in $\mathbb{R}^3$, and can be cut-and-flattened as a square $\mathbb{T}^2 = [-\pi, \pi) \times [-\pi, \pi)$ on $\mathbb{R}^2$ (cf. Figure 1, Eltzner, Huckemann and Mardia (2018)). Therefore, a point $(-\pi, -\pi)$ is closer to $(\pi - \epsilon, \pi - \epsilon)$ than $(-\pi + 2\epsilon, -\pi + 2\epsilon)$ for some $\epsilon > 0$. Due to this geometric constraint, as we shall see in Section 3.1, most off-the-shelf clustering methods are not applicable, at least not without a proper adaptation.

In this article we propose a novel approach for clustering on the torus, based on the conformal prediction framework (Lei, Robins and Wasserman (2013), Lei et al. (2018), Vovk, Gammerman and Shafer (2005)). The conformal prediction framework is a method of constructing distribution-free prediction sets with finite-sample validity, but there has been no attempt of applications to circular variables in the literature. Following Lei, Rinaldo and Wasserman (2015), Lei, Robins and Wasserman (2013), we construct estimators of prediction sets for the

Received September 2020; revised January 2021.
Key words and phrases. Directional statistics, torsion angles, von Mises distribution, prediction set, density estimation, protein structure.

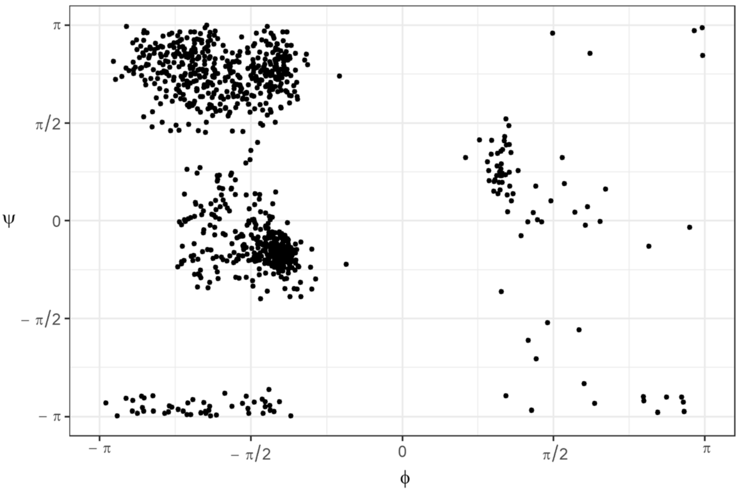


FIG. 1. *A Ramachandran plot or a* (ϕ, ψ)*-plot of the SADS-CoV protein structure* (*Yu et al.* (2020)).

protein torsion angles, based on a kernel density estimate on the torus (Di Marzio, Panzera and Taylor (2011)) and a finite mixture of bivariate von Mises distributions (Mardia, Taylor and Subramaniam (2007), Mardia et al. (2012)). The conformal prediction sets for toroidal data are described in Section 2; see Figure 2 for the prediction sets obtained from the data displayed in Figure 1. We construct several variants of prediction sets, based on which the proposed clustering of toroidal data is defined.

We propose to identify clusters by the connected components of prediction sets. We say $A_1, \ldots, A_K$ are connected components of (a closed set) $A \subset \mathbb{T}^2$ if A is the disjoint union

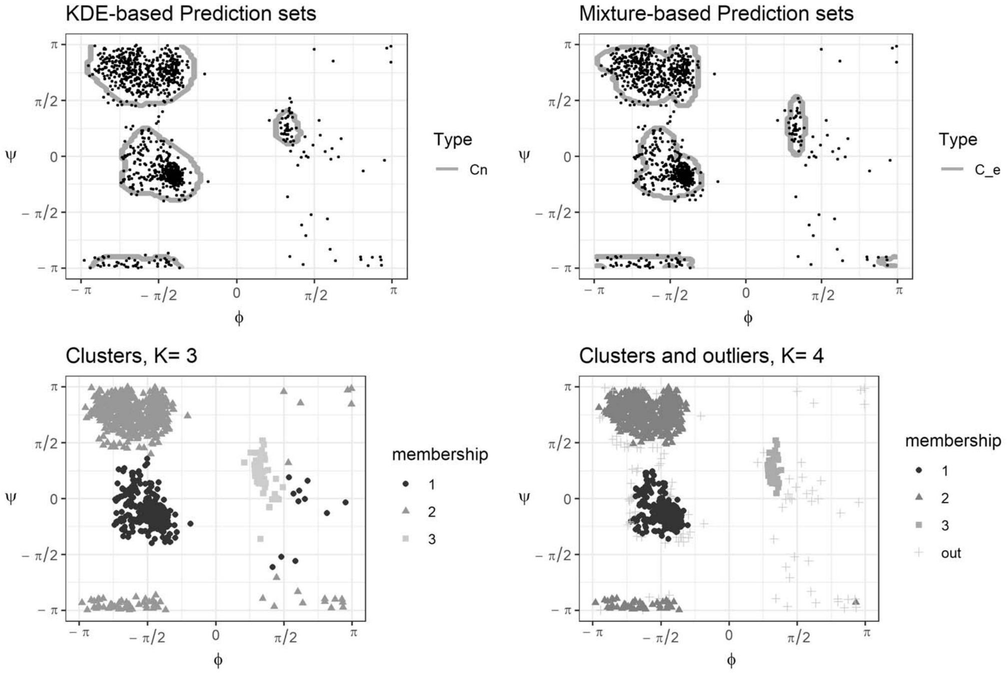


FIG. 2. *Top row*: *Prediction sets at level* 90% *are overlaid with the data displayed in Figure* 1; *see equations* (5) *and* (10) *in Section* 2. *Bottom row*: *Cluster memberships assigned by* A_e (*left*) *and* A_o (*right*); *see equations* (11) *and* (12) *in Section* 3.

of A_k's and each A_k cannot be divided into two disjoint nonempty closed sets. For a careful choice of conformity score (from which the prediction set $\hat{C}_n$ is estimated), the connected components are simply unions of ellipses, and an automated identification of clusters is possible. The resulting number K of clusters depends on the choice of the level of the prediction set and the hyperparameter used in fitting the mixture model. The problem of choosing K is then transformed into the problem of setting the level and the hyperparameter which are chosen by striking a balance between the desired coverage (the higher the better) and the volume of the prediction set (the smaller the better). We propose and compare two methods of cluster membership assignment; see Section 3 for the proposed approaches of clustering. Figure 2 demonstrates the results of the proposed clustering in which three major clusters are identified.

The data shown in Figures 1 and 2 are torsion angles from Cryo-EM structures of SADS-CoV spike glycoproteins, previously analyzed in Yu et al. (2020). The swine acute diarrhea syndrome coronavirus (SADS-CoV) is a strain of coronavirus and is known to be structurally similar to Rhinolophus bat coronavirus (HKU2) (Gong et al. (2017)). We compare the clusters of SADS-CoV torsion angles with those of HKU2 and the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a variant of RNA coronavirus (Walls et al. (2020)) which is the cause of the COVID-19 pandemic (Chan et al. (2020), Gorbalenya et al. (2020)). Section 4.1 is devoted to analyze the torsion angles from these coronavirus proteins, and we reveal a potential difference in the distributions of torsion angles of SARS-CoV-2 from those of SADS-CoV and HKU2.

The advantages of the proposed clustering framework are further highlighted in a simulation study (Section 4.2) by an empirical comparison with existing clustering methods on the torus. Our method takes the cyclic nature of toroidal data into account and performs superior when the clusters are of irregular shape. Technical details and supplementary figures are supplied in the Supplementary Material (Jung, Park and Kim (2021)).

The proposed clustering approaches can be easily adapted for the usual multivariate data (in the Euclidean space) and can be extended to other types of manifold-valued data in, for example, hyperspheres and higher-dimensional tori. In Section 5 we discuss related approaches in Euclidean spaces and point out future research directions.

1.1. *Tools to handle data on the torus.* Bivariate angular data can be understood as lying on the product of two unit circles and are naturally parameterized by angles in $\mathbb{T}^2 = [-\pi, \pi) \times [-\pi, \pi)$ or in $[0, 2\pi) \times [0, 2\pi)$. We use the former, but our discussion is invariant to the choice of parameterization.

Let $x = (\phi_x, \psi_x)$, $y = (\phi_y, \psi_y) \in \mathbb{T}^2$. The angular subtraction is defined as $x \ominus y = \arg(e^{i(x-y)}) := (\arg(e^{i(\phi_x - \phi_y)}), \arg(e^{i(\psi_x - \psi_y)}))$ in which operations are applied elementwise. (The range of the argument is $(-\pi, \pi]$.) As an example, $(\pi/2, 3\pi/4) \ominus (0, -3\pi/4) = (\pi/2, -\pi/2)$. A natural metric on the torus is

$$\rho(x, y) = \|x \ominus y\|_2 = \|y \ominus x\|_2 = \left[\left\{\arg\left(e^{i(\phi_x - \phi_y)}\right)\right\}^2 + \left\{\arg\left(e^{i(\psi_x - \psi_y)}\right)\right\}^2\right]^{1/2}.$$

For any $x, y \in \mathbb{T}^2$, $\rho(x, y) \in [0, \sqrt{2}\pi]$. The toroidal distance function $\rho(x, y)$ can be used, for example, in defining the nearest neighbors of a point on $\mathbb{T}^2$. Finally, given a set of points $\mathbb{X} = \{X_1, \dots, X_n\}$, the (sample) toroidal mean is defined by $m(\mathbb{X}) = \operatorname{argmin}_x \sum_{i=1}^n \rho^2(x, X_i)$, where the minimum is over $\mathbb{T}^2$. The toroidal mean always exists but may not be unique. Note that our parameterization of the torus with the metric ρ makes the torus a flat torus (cf. O'Neill (2006)).

2. Distribution-free prediction sets on the torus.

2.1. *Conformal prediction framework.* Suppose we observe an exchangeable sequence of random variables $\{X_1, \dots, X_n\}$ on the torus, where $X_i = (\phi_i, \psi_i) \in \mathbb{T}^2$. A level $1 - \alpha$

prediction set $C_n = C_n(X_1, \dots, X_n)$ satisfies, for new observation X_{n+1},

$$P(X_{n+1} \in C_n) \geq 1 - \alpha, \tag{1}$$

where the probability P is with respect to the exchangeable sequence $\{X_1, \dots, X_n, X_{n+1}\}$. Conformal prediction framework estimates C_n by introducing *conformity scores*. The conformity score $\sigma_i = \sigma(\mathbb{X}_{n+1}; X_i)$ measures the degree of conformity of X_i, compared to the augmented sample $\mathbb{X}_{n+1} = \{X_1, \dots, X_n, X_{n+1}\}$, and is defined for each X_i, $i = 1, \dots, n+1$. For example, a conformity score defined by $\sigma_i := \sqrt{2}\pi - \rho(m(\mathbb{X}_{n+1}), X_i)$ for the toroidal mean $m(\mathbb{X}_{n+1})$, is higher (i.e., more conformal) as X_i being closer to the mean. Since X_{n+1} is not observed, the conformity score σ_i is computed for each and every candidate $x \in \mathbb{T}^2$ by setting $X_{n+1} = x$. The prediction set given by the conformal prediction framework is

$$C_n = \{x \in \mathbb{T}^2 : \xi_n(x) > \alpha\}, \tag{2}$$

where

$$\xi_n(x) = \frac{1}{n+1} \sum_{j=1}^{n+1} \mathbf{1}_{\sigma_j \leq \sigma_{n+1}}.$$

It can be shown that, for any choice of conformity score σ, the conformal prediction set C_n is a valid level $1 - \alpha$ prediction set, that is, (1) holds if $\mathbb{X}_{n+1}$ is exchangeable; see, for example, page 26 of Vovk, Gammerman and Shafer (2005) and Section 2 of Lei, Robins and Wasserman (2013). A caveat here is that, for a poorly-chosen conformity score σ, the volume of C_n can be nontrivially larger than using more sensible choices of σ. Lei, Robins and Wasserman (2013) have shown that if kernel density estimates are used for conformity scores, then C_n is efficient for data in $\mathbb{R}^d$. In the next subsection we adapt the idea of Lei, Robins and Wasserman (2013) for the prediction of toroidal data.

2.2. *Prediction set by kernel density estimates on the torus.* A natural population counterpart of prediction set C_n is the density level set $L(t) = \{x \in \mathbb{T}^2 : p(x) \geq t\}$, where $p(\cdot)$ is the density function of X_{n+1} on $\mathbb{T}^2$. For a given α, $t = t_\alpha$ is defined as the largest t satisfying $P(X_{n+1} \in L(t)) \geq 1 - \alpha$. If the density function p is continuous on $\mathbb{T}^2$ and is not flat at $\{x : p(x) = t_\alpha\}$, then it can be shown that the level set $L(t_\alpha)$ has exact coverage and has the smallest volume among all level $1 - \alpha$ prediction sets (Polonik (1997)). To be precise, let $\mu(A)$ be the area of a subset $A \subset \mathbb{T}^2$, scaled by $(2\pi)^2$, so that $\mu(\mathbb{T}^2) = 1$. Then, $L(t_\alpha) = \arg\min_C \mu(C)$ where the minimum is over $\{C \subset \mathbb{T}^2 : P(X_{n+1} \in C) \geq 1 - \alpha\}$.

We build a conformal prediction set C_n using a form similar to the density level set $L(t_\alpha)$ in which p and t_α are given by a kernel density estimate (kde) $\hat{p}$ and conformity scores; see (5) below. This is done indirectly by using $\hat{p}(\cdot)$ as the conformity score $\sigma(\mathbb{X}_{n+1}, \cdot)$.

We first discuss a kde on $\mathbb{T}^2$. Due to the doubly cyclic nature of the sample space $\mathbb{T}^2$, care is needed in defining a kde. For example, scalable kernels, defined on a bounded domain such as the Epanechnikov kernel $K(x) = \frac{3}{4}(1 - x^2)\mathbf{1}_{|x| \leq 1}$ (or any generalization to $\mathbb{T}^2$), do not integrate to 1 when scaled by $x \mapsto x/s$, $s > \pi$. Guassian kernels suffer from the need of truncation on $[-\pi, \pi]$ or of wrapping around $\mathbb{T}^2$. Instead, following Di Marzio, Panzera and Taylor (2011), we use a two-product von Mises kernel (with a common concentration parameter κ): For $x = (\phi, \psi) \in \mathbb{T}^2$,

$$K_\kappa(x) = \frac{e^{\kappa \cos(\phi)}}{2\pi I_0(\kappa)} \frac{e^{\kappa \cos(\psi)}}{2\pi I_0(\kappa)}, \tag{3}$$

where I_ν is the modified Bessel function of the first kind of order ν. The von Mises density $(2\pi I_0(\kappa))^{-1} e^{\kappa \cos(\phi)}$ plays the role of the normal distribution for circular statistics (Mardia and Jupp (2000)). The kernel (3) is a simple form of toroidal kernels. Kernel density estimates, using toroidal kernels, are known to enjoy some nice asymptotic properties (Di Marzio, Panzera and Taylor (2011)). Using (3), the kde of $p(u)$ at $u \in \mathbb{T}^2$ is $\hat{p}_n(u) = \sum_{i=1}^n K_\kappa(u - X_i)/n$, where the cyclic nature of $\mathbb{T}^2$ is handled through the cosine function.

To define conformity scores, we use the kde based on the augmented data $\mathbb{X}_{n+1}$,

$$\begin{aligned} \hat{p}_{n+1}(u) &= \frac{1}{n+1}\sum_{i=1}^{n+1} K_\kappa(u - X_i) \\ &= \frac{1}{n+1}\sum_{i=1}^{n} K_\kappa(u - X_i) + \frac{1}{n+1} K_\kappa(u - x) \end{aligned} \tag{4}$$

in which the unobserved X_{n+1} is replaced by $X_{n+1} = x$ for an inspection point $x \in \mathbb{T}^2$. Setting $\sigma_i = \hat{p}_{n+1}(X_i)$, the expression $\xi_n(x)$ in the conformal prediction set (2) becomes $\xi_n(x) = \frac{1}{n+1}\{1 + \sum_{i=1}^n \mathbf{1}_{\hat{p}_{n+1}(X_i) \le \hat{p}_{n+1}(x)}\}$. Observe that $\xi_n(x) \in \{\frac{1}{n+1}, \dots, \frac{n}{n+1}, 1\}$, which in turn leads that $\pi_n(x) > \alpha$ is equivalent to $\pi_n(x) > \tilde{\alpha}$, where

$$\tilde{\alpha} = \frac{i_{n,\alpha}}{n+1}, \quad i_{n,\alpha} = \lfloor (n+1)\alpha \rfloor.$$

To further simplify, let $X_{(i)}$ $(i = 1, \dots, n)$ satisfy $\hat{p}_{n+1}(X_{(i)}) \le \hat{p}_{n+1}(X_{(i+1)})$ for $1 \le i \le n-1$. Then,

$$\begin{aligned} C_n &= \{x \in \mathbb{T}^2 : \xi_n(x) > \tilde{\alpha}\} \\ &= \{x \in \mathbb{T}^2 : \hat{p}_{n+1}(x) \ge \hat{p}_{n+1}(X_{(i_{n,\alpha})})\}. \end{aligned} \tag{5}$$

The equation (5) is verified in Appendix B.1 in the Supplementary Material (Jung, Park and Kim (2021)). Note that C_n is not exactly a level set of a single function, as the density estimate $\hat{p}_{n+1}$ depends on x. Moreover, to evaluate whether $x \in C_n$ for many inspection points x, the computation of $\hat{p}_{n+1}$ and $X_{(i)}$ is required for each x. We provide two approximations of (5) as level sets of $\hat{p}_n$, which require a significantly less computation, in Appendix B.2 in the Supplementary Material (Jung, Park and Kim (2021)).

As a demonstration, the prediction set C_n of level 0.9 is plotted in Figure 2 (top left panel).

2.3. *Prediction set by mixtures of bivariate von Mises.* In this section a mixture density and its variants are used as a conformity score σ in constructing prediction sets.

Our choice of the mixture model on the torus is composed of bivariate von Mises distributions (Chakraborty and Wong (2017), Mardia, Taylor and Subramaniam (2007)). In particular, we use the sine variant of bivariate von Mises density (Singh, Hnizdo and Demchuk (2002)),

$$f(x) = C \exp[\kappa_1 \cos(\phi - \mu_1) + \kappa_2 \cos(\psi - \mu_2) + \lambda \sin(\phi - \mu_1)\sin(\psi - \mu_2)],$$

for $x = (\phi, \psi) \in \mathbb{T}^2$, where $(\mu_1, \mu_2) \in \mathbb{T}^2$ is the location parameter, $\kappa_1 > 0$, $\kappa_2 > 0$ are concentration parameters and λ, satisfying $\lambda^2 < \kappa_1\kappa_2$, determines an association of two circular variables. The normalizing constant is given by

$$C^{-1} = (2\pi)^2 \sum_{m=0}^{\infty} \binom{2m}{m} \left(\frac{\lambda^2}{4\kappa_1\kappa_2}\right)^m I_m(\kappa_1) I_m(\kappa_2). \tag{6}$$

If $\lambda = 0$, then $f(\phi, \psi)$ is a product of two von Mises densities. For large concentrations (larger values of κ_1, κ_2), the density is well approximated by a bivariate normal density. For simplicity, assume $\mu_1 = \mu_2 = 0$ for now. Then,

$$f(x) \approx C' \exp\left(-\frac{1}{2} x^T \Sigma^{-1} x\right), \quad \Sigma^{-1} = \begin{pmatrix} \kappa_1 & -\lambda \\ -\lambda & \kappa_2 \end{pmatrix}, \tag{7}$$

where $C' = |2\pi\Sigma|^{-1/2} = \sqrt{\kappa_1\kappa_2 - \lambda^2}/(2\pi)$. The parameters κ_1, κ_2 and λ can be interpreted using (7): Writing σ_ϕ^2, σ_ψ^2 and ρ for the variances of ϕ and ψ and the correlation coefficient between them, respectively, we have $\kappa_1^{-1} \approx \sigma_\phi^2(1-\rho^2)$, $\kappa_2^{-1} \approx \sigma_\psi^2(1-\rho^2)$ and $\lambda \approx \rho/\{\sigma_\phi\sigma_\psi(1-\rho^2)\}$. We require $\lambda \in (-\sqrt{\kappa_1\kappa_2}, \sqrt{\kappa_1\kappa_2})$, under which condition the bivariate von Mises density is unimodal (Theorem 3, Mardia, Taylor and Subramaniam (2007)).

For a $J = 1, 2, \dots$, a J-mixture density for the toroidal data is given by $p(x) = \sum_{j=1}^J \pi_j f_j(x)$, where π_j's are mixing probabilities ($\sum_j \pi_j = 1$) and $f_j(\cdot)$ is the bivariate von Mises density with parameters $\theta_j := (\mu_{1j}, \mu_{2j}, \kappa_{1j}, \kappa_{2j}, \lambda_j)$.

Maximum likelihood estimates of π_j, θ_j and $f_j(\cdot)$, based on $\mathbb{X}_n = \{X_1, \dots, X_n\}$, are denoted by $\hat{\pi}_j$, $\hat{\theta}_j$ and $\hat{f}_j(\cdot)$, respectively, which provide a natural definition of $\hat{p}_n^{\mathrm{M}}(\cdot)$ (M stands for mixture models). These estimates are numerically obtained by an EM algorithm. The details of the estimation procedure are discussed in Appendix A.1 in the Supplementary Material (Jung, Park and Kim (2021)). Let $\hat{p}_{n+1}^{\mathrm{M}}$ be defined similarly for an augmented data set $\mathbb{X}_{n+1}$. Then,

$$C_n^{\mathrm{M}} = \{x \in \mathbb{T}^2 : \hat{p}_{n+1}^{\mathrm{M}}(x) \geq \hat{p}_{n+1}^{\mathrm{M}}(X_{(i_{n,\alpha})})\}$$

is a valid level $1-\alpha$ prediction set constructed by setting $\sigma_i = \hat{p}_{n+1}^{\mathrm{M}}(X_i)$; compare with equation (5) and the preceding discussion.

Computing $\hat{p}_{n+1}^{\mathrm{M}}$ for many x values of the augmented data $\mathbb{X}_{n+1}$ with $X_{n+1} = x$ is not practical. Instead, we use a sample splitting strategy, called *inductive conformal prediction* in the literature of conformal prediction (Lei, Rinaldo and Wasserman (2015), Vovk, Gammerman and Shafer (2005)). The idea is to randomly split $\mathbb{X}_n$ into $\mathbb{X}_{(1)}$ of size n_1 and $\mathbb{X}_{(2)}$ of size n_2 ($n_1 + n_2 = n$). The first part $\mathbb{X}_{(1)}$ is used in the estimation of p, resulting in $\hat{p}_{n_1}^{\mathrm{M}}$ (denoted by $\hat{p}$ hereafter for notational simplicity). The conformity scores are computed for the second part $\mathbb{X}_{(2)} = \{X_{n_1+1}, \dots, X_n\}$: $\sigma_i = \hat{p}(X_{n_1+i})$ ($i = 1, \dots, n_2$), then sorted to $\sigma_{(1)} \leq \cdots \leq \sigma_{(n_2)}$. A level $1-\alpha$ inductive conformal prediction set is the level set of $\hat{p}$,

$$\hat{C}_n = \{x \in \mathbb{T}^2 : \hat{p}(x) \geq \sigma_{(i_{n_2,\alpha})}\}, \quad i_{n_2,\alpha} = \lfloor (n_2+1)\alpha \rfloor. \tag{8}$$

As the computation of $\hat{p}$ is required only once in (8), evaluating the inductive prediction set has a huge computational advantage over evaluating C_n^{M}. It is shown in Lei, Rinaldo and Wasserman (2015) that inductive conformal prediction set also enjoys the distribution-free finite-sample validity, that is, $P(X_{n+1} \in \hat{C}_n) \geq 1-\alpha$. We emphasize that $\hat{C}_n$ is valid not only when the mixture density estimate $\hat{p}_{n_1}^{\mathrm{M}}$ is used in σ_i but also for *any* choices of conformity scores. Thus, the variants of $\hat{C}_n$ we discuss below also have the finite-sample validity.

Suppose that the components in $\hat{p}$ are well separated, then

$$\hat{p}(x) = \sum_{j=1}^J \hat{\pi}_j \hat{f}_j(x) \approx \hat{p}^{\max}(x) = \max_j \hat{\pi}_j \hat{f}_j(x).$$

This suggests setting the conformity scores to $\sigma_i = \hat{p}^{\max}(X_{n_1+i})$ which, in turn, leads to a valid inductive conformal prediction set (even if the components are not separated).

Another variant is obtained when considering the normal approximation to the bivariate von Mises (7). The log-max-density $\log(\hat{p}^{\max}(x)) = \max_j\{\log(\hat{\pi}_j) + \log(\hat{f}_j(x))\}$ is approximated by $\hat{e}(x)/2 + \text{constant}$, where

$$\begin{aligned} &\hat{e}(x) = \max_j \hat{e}_j(x), \\ &\hat{e}_j(x) = -(x \ominus \hat{\mu}_{(j)})^T \widehat{\Sigma}_j^{-1}(x \ominus \mu_{(j)}) + \log\bigl((\hat{\kappa}_{1j}\hat{\kappa}_{2j} - \hat{\lambda}_j^2)\hat{\pi}_j^2\bigr). \end{aligned} \tag{9}$$

Here, $\hat{\mu}_{(j)}^T = (\hat{\mu}_{1j}, \hat{\mu}_{2j})$, and $\widehat{\Sigma}_j^{-1}$ is given in (7) with the parameters replaced by the jth component estimates. The "$x \ominus y$" notation refers to the angular subtraction, defined in Section 1.1. Any level set of $\hat{e}(\cdot)$ is a union of ellipses on $\mathbb{T}^2$, as shown in Lemma 2.1.

LEMMA 2.1. *Let $\hat{e}(x)$ be defined in* (9), *and let*

$$c_j = \log\big((\hat{\kappa}_{1j}\hat{\kappa}_{2j} - \hat{\lambda}_j^2)\hat{\pi}_j^2\big)$$

and

$$\hat{E}_j(t) = \big\{x \in \mathbb{T}^2 : (x \ominus \hat{\mu}_{(j)})^T \widehat{\Sigma}_j^{-1}(x \ominus \mu_{(j)}) \leq c_j - t\big\}.$$

Then, for any $t \in \mathbb{R}$, $L^{\hat{e}}(t) := \{x \in \mathbb{T}^2 : \hat{e}(x) \geq t\} = \bigcup_{j=1}^{J} \hat{E}_j(t)$.

PROOF OF LEMMA 2.1. Write $\hat{e}(x) = \max_j \hat{e}_j(x)$, and observe that $\hat{e}_j(x) \geq t$ is equivalent to the condition for $x \in \hat{E}_j(t)$. Then, $\{x \in \mathbb{T}^2 : \max_j \hat{e}_j(x) \geq t\} = \{x \in \mathbb{T}^2 : \hat{e}_j(x) \geq t \text{ for some } j\} = \bigcup_j \{x \in \mathbb{T}^2 : \hat{e}_j(x) \geq t\} = \bigcup_j \hat{E}_j(x)$, as required. □

Each $\hat{E}_j(t)$ is an ellipse,[1] provided that $t < c_j$. Otherwise, $\hat{E}_j(t) = \varnothing$ for $c_j - t \leq 0$. Note that as t increases, the ellipses $\hat{E}_j(t)$ of $L^{\hat{e}}(t)$ become simultaneously smaller, and the number of ellipses involved decreases (as some $\hat{E}_j(t)$ degenerates to an empty set).

Setting $\sigma_i = \hat{e}(X_{n_1+i})$ (which are then ordered to satisfy $\sigma_{(1)} \leq \cdots \leq \sigma_{(n_2)}$), the inductive conformal prediction set is

$$\hat{C}_n^e = \big\{x \in \mathbb{T}^2 : \hat{e}(x) \geq \sigma_{(i_{n_2,\alpha})}\big\} = L^{\hat{e}}(\sigma_{(i_{n_2,\alpha})}). \tag{10}$$

An advantage of using $\hat{C}_n^e$ is its simple form (the union of ellipses), which is computationally handy, when we consider clustering in Section 3.

We denote the inductive prediction sets by $\hat{C}_n^{\text{mix}}$ if $\sigma(\cdot) = \hat{p}(\cdot)$ (8), $\hat{C}_n^{\max}$ if $\sigma(\cdot) = \hat{p}^{\max}(\cdot)$, and $\hat{C}_n^e$ if $\sigma(\cdot) = \hat{e}(\cdot)$ (10). The prediction set $\hat{C}_n^e$ is plotted in Figure 2 for the SADS-CoV torsion angles, further discussed in Section 4.1. (All prediction sets are displayed in the Appendix Figure C.2 in the Supplementary Material (Jung, Park and Kim (2021)).) An inductive prediction set $\hat{C}_n^{\text{kde}}$, given by the kde, is defined similarly.

3. Clustering on the torus.

3.1. *Clustering on the torus*: *An overview.* Due to the cyclic nature of data on $\mathbb{T}^2$, recklessly applying off-the-shelf clustering methods can result in unstable and low-quality clustering results. For example, the result of a naive k-means clustering applied to $x_i \in \mathbb{T}^2$ is different from the clustering results from the data $x_i' := x_i + (\pi, \pi) \in [0, 2\pi)^2$. However, since the data set $\{x_i'\}$ is simply translated from $\{x_i\}$, the clustering result should be identical. Moreover, an apparent cluster near the border of the square $[-\pi, \pi) \times [-\pi, \pi)$ is split into two or more clusters, as exemplified in the top left panel of Figure 3. Some existing work on clustering toroidal data (Kountouris and Hirst (2009)) suffers from the same problem.

We briefly review adaptations of a few popular clustering methods for use on the torus.

The k-means clustering is by far the most popular off-the-shelf clustering algorithm, obtained by recursively partitioning observations and computing cluster centers. A straightforward adaptation of the k-means for data on the torus is given by using the toroidal distance $\rho(x, c)$ (in partitioning) and the toroidal mean $m(\mathbb{X})$ (for cluster centers). An approximation of this *intrinsic* k-means algorithm was considered in Gao et al. (2018). In contrast, an *extrinsic* k-means algorithm uses the ambient space for $\mathbb{T}^2$ in which each

[1] Precisely, the set $\{x \in \mathbb{T}^2 : (x \ominus \mu)^T S^{-1}(x \ominus \mu) \leq 1\}$ is an ellipse for small enough S; for large S, the set is the intersection of an ellipse with the square of width 2π centered at μ. In our analysis, $S = \max\{c_j - t, 0\}\hat{\Sigma}_j$ is typically small.

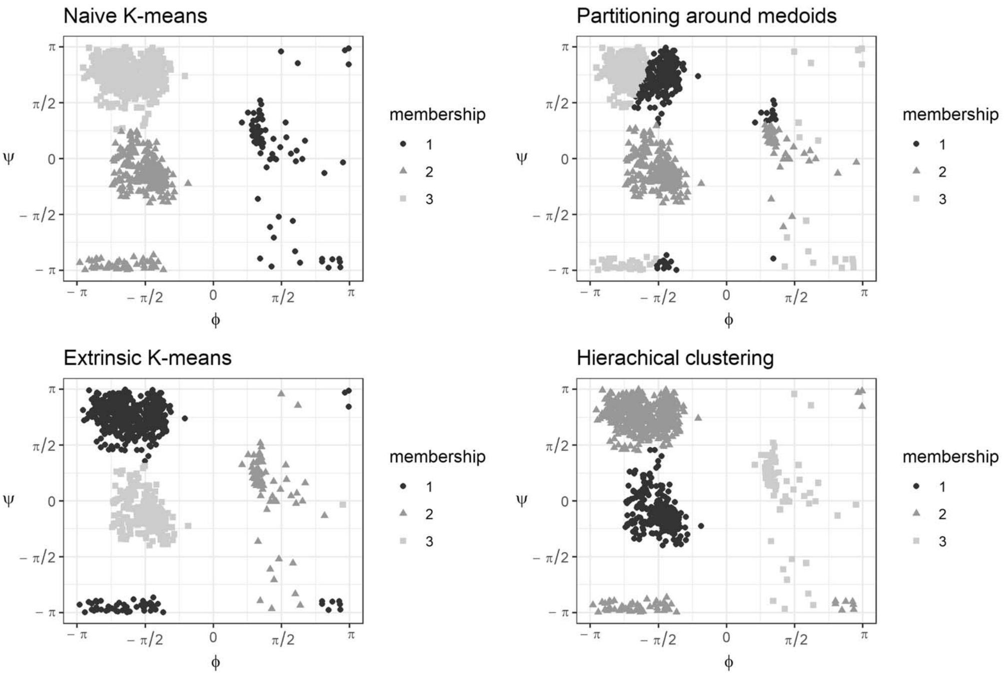


FIG. 3. *Existing methods for clustering on the torus.*

$\mathbb{T} = [-\pi, \pi)$ is embedded as the unit circle in $\mathbb{R}^2$. Each $x_i = (\phi_i, \psi_i) \in \mathbb{T}^2$ is transformed to $x_i' = (\cos\phi_i, \sin\phi_i, \cos\psi_i, \sin\psi_i) \in \mathbb{R}^4$, then a usual k-means algorithm is applied to x_i''s. This extrinsic k-means algorithm is computationally advantageous, as it uses the vector operations in $\mathbb{R}^4$, and any variant of the k-means algorithm can be used as well. In Figure 3 we demonstrate the result of the extrinsic k-means clustering for the protein structure data with initial values given by the k-means++ algorithm (Arthur and Vassilvitskii (2007)).

A number of clustering algorithms depend only on the pairwise distances (or similarity measures) of data (Xu and Tian (2015)). The partitioning around medoids (PAM) algorithm of Kaufman and Rousseeuw (2009) [originally proposed in 1990], its variants (van der Laan, Pollard and Bryan (2003)) and hierarchical clustering methods (Murtagh and Contreras (2012), Murtagh and Contreras (2017)) are prominent examples. All of these clustering methods are readily available for data on the torus with pairwise toroidal distances $\{\rho(x_i, x_j) : 1 \le i < j \le n\}$ as an input.

Gaussian mixture models are another popular and well-developed class of clustering methods for vector-valued data; see, for example, Scrucca et al. (2016). A toroidal adaptation of mixture models is given in Section 2.3 in which we used bivariate von Mises mixture models (Mardia, Taylor and Subramaniam (2007)). Using a fitted mixture model $\hat{p}(\cdot) = \sum_j \hat{\pi}_j \hat{f}_j(\cdot)$, a soft clustering of an inspection point x is the probability of the cluster membership Y, $\hat{P}(Y = j \mid X = x) = \hat{\pi}_j \hat{f}_j(x)/\hat{p}(x)$ for $j = 1, \dots, K$. A hard clustering of x is simply $\arg\max_j \hat{\pi}_j \hat{f}_j(x)$.

Examples in Figure 3 suggest that, when the apparent clusters are of irregular shapes (rather than having elliptical shapes), all of the methods above provide unsatisfactory clustering results. For such cases a density level set clustering might be useful (Cheng (1995), Hartigan (1975)). For a multimodal density p, the level set $L(t) = \{x \in \mathbb{T}^2 : p(x) \ge t\}$ for appropriately chosen t separates each mode from the others. However, it is unclear what should be an actual clustering function that assigns a cluster label for each $x \in \mathbb{T}^2$. A naive approach of *partitioning around modes* works poorly when the clusters (or the connected components of $L(t)$) are of irregular shapes.

In the next subsection we develop a conformity-score level set clustering for the torus and show that a sensible and automated cluster-label assignment can be done by employing a carefully chosen conformity score.

3.2. *Clustering by conformal prediction sets.* A conformal prediction set C_n on the torus developed in Section 2 provides a natural clustering of $\mathbb{T}^2$, where a cluster is given by a connected component of C_n. If an inductive conformal prediction set (8) or (10) is used instead of C_n, then the corresponding clustering is equivalent to a density level set clustering (Hartigan (1975)), except that the density function $p(x)$ is replaced by the conformity score $\sigma(x)$ which is not necessarily a density estimate.

We point out that a prediction set (denoted by $\hat{C}$, representing either C_n (5), $\hat{C}_n^{\text{mix}}$ (8), $\hat{C}_n^e$ (10) or $\hat{C}_n^{\text{kde}}$) depends on the choice of level $1-\alpha$ and a hyperparameter (concentration κ for a kde-based prediction set or the number J of mixture components for a mixture-model-based prediction set). The inherent problem of choosing the number K of clusters is substituted by the problem of selecting α and the hyperparameter. Our approach for the selection of level and hyperparameter will be discussed in Section 3.3. For now, let us assume that a prediction set is given for a prespecified choice of α and κ (or J).

Identification of connected components. For any given prediction set $\hat{C}$, its connected components and the number of distinct components K can be computed algorithmically using a fine grid on $\mathbb{T}^2$. Our visual illustrations of $\hat{C}$ in Figure 2 (top row) are indeed given by evaluating $\mathbf{1}_{x\in\hat{C}}$ for $x \in T^2$, where $T = \{\pi(2t-1) : t = 1/100, 2/100, \dots, 1\}$. To identify all grid points of a connected component containing x, one may use a flood fill algorithm, that is, recursively identifying neighborhoods of x in $\hat{C}$, which is then applied to all unidentified points in $\hat{C}$ to assign x with labels distinct for each connected component. A drawback of such an algorithmic approach is that its accuracy deteriorates when the grid T is coarse.

When the elliptical prediction set $\hat{C}_n^e$ is used, an exact identification of the connected components is possible. Recall that $\hat{C}_n^e = \bigcup_{j=1}^J \hat{E}_j$, where $\hat{E}_j = \hat{E}_j(\sigma_{(i_{n_2,\alpha})})$, as defined in Lemma 2.1. In this case the connected components are exactly unions of ellipses. To identify the connected components, we create an adjacent matrix A (of size $J \times J$) whose (i,j)th element is 1 if $\hat{E}_i \cap \hat{E}_j \neq \varnothing$, 0 otherwise. (We discuss complications in testing whether two toroidal ellipses intersect in Appendix A.2 in the Supplementary Material (Jung, Park and Kim (2021)).) The adjacent matrix A gives rise to an undirected graph (where nodes are labeled $1,\dots,J$) whose connected components are easily found by a simple breadth first search. If $\hat{E}_j = \varnothing$, then the corresponding node j is removed from the graph. Denoting the connected components of the graph by the node indices $\ell_1,\dots,\ell_K \subset \{1,\dots,J\}$, the clusters (connected components) of $\hat{C}_n^e$ are

$$\mathcal{E}_k = \bigcup_{j\in\ell_k} \hat{E}_j, \quad k = 1,\dots,K.$$

Note that $\mathcal{E}_k \cap \mathcal{E}_{k'} = \varnothing$ for $k \neq k'$, $\hat{C}_n^e = \bigcup_{k=1}^K \mathcal{E}_k$, and the number of clusters is K.

As an illustration, Figure 4 displays a toy data set on $\mathbb{T}^2$, overlaid with $\hat{E}_j$'s; the union of which provides a 90% prediction set. For this data set, $\mathcal{E}_1 = \hat{E}_1$, $\mathcal{E}_2 = \hat{E}_2 \cup \hat{E}_3$ and there are $K = 2$ clusters.

Cluster assignment. Every $x \in \hat{C}$ has a natural cluster membership assignment, given by the label of a connected component it belongs to.

For $x \notin \hat{C}$, a natural approach of membership assignment is to find the "closest" cluster. When the closeness is defined by the toroidal distance between x and a connected component $\mathcal{E}_k$, finding the closest cluster can be handled by a grassfire transformation (Blum (1967)).

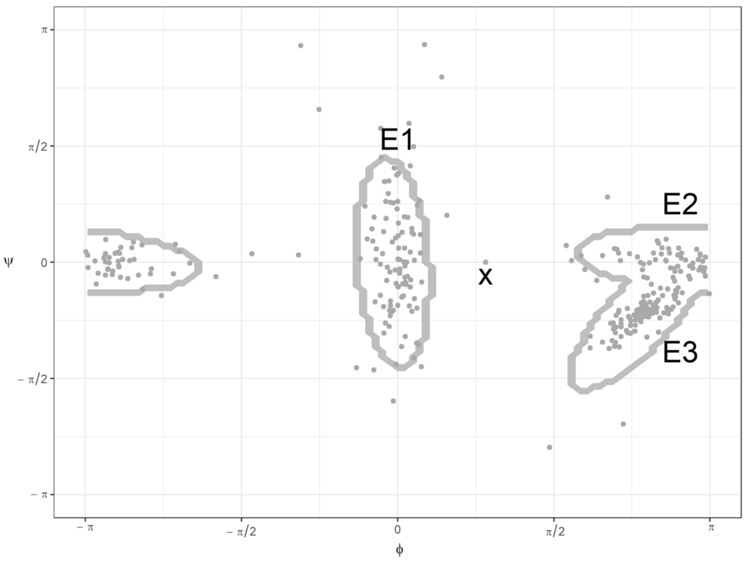


FIG. 4. *Clustering by the conformal prediction set $\hat{C}_n^e$. The prediction set $\hat{C}_n^e$ is obtained by a four-mixture of bivariate von Mises with level $1-\alpha = 90\%$. Here, $\hat{E}_4 = \varnothing$ is not shown. There are $K = 2$ connected components, and the point x is assigned to the second cluster $\mathcal{E}_2$ by $A_e(x) = 2$* (11).

However, this approach has a statistical drawback. To motivate, consider two components of a bivariate von Mises (vM2) mixture: One is approximately given by a vM2 with parameters $\theta_1 = (0, 0, 50, 3, 0)$, the other by $\theta_2 = (-\pi, 0, 3, 50, 0)$. Ellipses corresponding to these two vM2 components are shown in Figure 4. There, x is closer to $\hat{E}_1$ in terms of the toroidal distance but is closer to $\hat{E}_2$ in terms of a Mahalanobis distance. Taking the cluster size into account, we propose a membership assignment rule based on the probability of cluster membership at x.

Suppose that the clustering is based on the mixture model fitting and that X follows the J-mixture of bivariate von Mises distributions, and let $Y \in \{1, \ldots, J\}$ be the unobservable component membership of X. Given connected component indices $\ell_1, \ldots, \ell_K \subset \{1, \ldots, J\}$, an $x \in \mathbb{T}^2$ is assigned to cluster $\hat{y}(x) = \arg\max_k \hat{P}(Y \in \ell_k | X = x)$, where $\hat{P}(Y \in \ell_k | X = x)$ is proportional to $\sum_{j \in \ell_k} \hat{\pi}_j \hat{f}_j(x)$. Denote this assignment rule as A_p so that, for any $x \in \mathbb{T}^2$,

$$A_p(x) = \arg\max_k \sum_{j \in \ell_k} \hat{\pi}_j \hat{f}_j(x).$$

If components are well separated, then $\sum_{j \in \ell_k} \hat{\pi}_j \hat{f}_j(x) \approx \max_{j \in \ell_k} \hat{\pi}_j \hat{f}_j(x)$ and the maximum over k is well approximated by the maximum over k of $\max_{j \in \ell_k} \hat{e}_j(x)$; see (9). Thus, an alternative cluster assignment rule, especially for use with $\hat{C}_n^e$, is given by assigning the cluster label of x to k if $\arg\max_j \hat{e}_j(x) \in \ell_k$,

$$A_e(x) = k \quad \text{if } \arg\max_j \hat{e}_j(x) \in \ell_k, k = 1, \ldots, K. \tag{11}$$

Note that for the mixture-based $\hat{C}_n^{\text{mix}}$ (8), the assignment rule A_p guarantees that, for $x \in \hat{C}_n^{\text{mix}}$, $A_p(x)$ is the label of a connected component to which it belongs. Likewise, for the ellipse-based $\hat{C}_n^e$, $A_e(x)$ has the correct label.

Another approach is simply creating a new label, representing outliers, for all $x \notin \hat{C}$,

$$A_o(x) = \begin{cases} k, & x \in \mathcal{E}_k, k = 1, \ldots, K; \\ \text{“outliers,”} & \text{otherwise.} \end{cases} \tag{12}$$

3.3. *Selection of level, hyperparameters and the number of clusters.* Our proposed clustering, introduced in Section 3.2, depends on the choice of level $1-\alpha$ and a hyperparameter (the number of mixture components J if using $\hat{C}_n^e$ or the concentration parameter κ if using $\hat{C}_n^{\text{kde}}$). Without losing generality, we focus on the clustering based on $\hat{C}_n^e$ and present a data-driven approach of choosing α and J. Write $\hat{C}_n^e = \hat{C}_n^e(\alpha, J)$, as it depends on α and J. The number of clusters K is just the number of connected components of $\hat{C}_n^e(\alpha, J)$.

As a prediction set, a desired level $1-\alpha$ of $\hat{C}_n^e$ may be predefined. For a fixed α, which choice of J provides the best prediction set? This question was answered for a kde-based prediction set in Lei, Robins and Wasserman (2013). Since the coverage probability of $\hat{C}_n^e(\alpha, J)$ is guaranteed to exceed $1-\alpha$, a prediction set with the smallest volume is desirable. The optimal choice of J, given an α, is

$$J_\alpha = \arg\min_J \mu\big(\hat{C}_n^e(\alpha, J)\big), \tag{13}$$

where the area $\mu(A)$ of $A \subset \mathbb{T}^2$ is scaled to satisfy $\mu(\mathbb{T}^2) = 1$.

With clustering in mind, there is no predetermined level $1-\alpha$, but the number J of mixture components or, equivalently, the fitted J-mixture bivariate von Mises model $(\hat{\pi}_j, \hat{f}_j(\cdot))$ may be given. The choice of α affects the number of clusters K more gravely than the choice of J. As the coverage $1-\alpha$ becomes larger, more components are connected to each other which results in a smaller number of clusters. Ideally, one would choose as large coverage $1-\alpha$ as possible and, simultaneously, wish for as small volume of the prediction set $\hat{C}_n^e(\alpha, J)$ as possible. However, there is a trade-off between large coverage (or, equivalently, small α) and small $\mu(\hat{C}_n^e(\alpha, J))$. We propose to choose α by

$$\alpha_J = \arg\min_\alpha \alpha + \mu\big(\hat{C}_n^e(\alpha, J)\big). \tag{14}$$

Choosing α by α_J is equivalent to finding an "elbow" of the graph of the function $\alpha \mapsto \mu(\hat{C}_n^e(\alpha, J))$. As an attempt of interpretation, we note that increased coverage (by choosing $\alpha < \alpha_J$) leads for the volume of prediction sets to increase sharply; on the other hand, by reducing the volume ($\alpha > \alpha_J$), the coverage probability $1-\alpha$ decreases fast. Setting the coverage at $1-\alpha_J$ strikes a balance.

To choose both α and J altogether, (13) and (14) are combined to

$$(\hat{\alpha}, \hat{J}) = \arg\min_{\alpha, J} \alpha + \mu\big(\hat{C}_n^e(\alpha, J)\big) \tag{15}$$

which is equivalent to finding the most lower-left point of $\{(\alpha, \mu(\hat{C}_n^e(\alpha, J))) : \alpha \in (0, 1), J = 1, \dots, J_{\max}\}$.

4. Clustering protein torsion angles.

4.1. *Analysis of torsion angles from coronavirus protein structures.* We analyze torsion angles from Cryo-EM structures of SADS-CoV, HKU2 and SARS-CoV-2 (coronavius) spike glycoproteins (Walls et al. (2020), Yu et al. (2020)). Coronaviruses are a large group of viral pathogens and pose severe threats to world healths when transmitted to humans. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a single-stranded RNA coronavirus, and is contagious in humans (Chan et al. (2020), Gorbalenya et al. (2020)). The swine acute diarrhea syndrome coronavirus (SADS-CoV), widespread in 2017 among commercial pigs in China, is found to share identical sequences up to 95% with Rhinolophus bat coronavirus (HKU2) (Gong et al. (2017)). In this analysis we confirm that SADS-CoV and HKU2 have almost identical protein backbone torsion angles, but SARS-CoV-2 has different sets of torsion angle clusters when compared to SADS-CoV.

We use the (ϕ, ψ) torsion angles extracted from SADS-CoV to illustrate our approaches of prediction sets and clustering and use the respective torsion angle data from HKU2 and SARS-CoV-2 as a validation set. The clustering of torsion angles from SARS-CoV-2 turns out to be different from the clustering for SADS-CoV. The data are available to public in the worldwide Protein Data Bank (PDB) with codes 6VXX, 6M15 and 6M16. The R packages `bio3d` (Grant et al. (2006)) was used to extract the data from the PDB.

The SADS-CoV protein structure is available from sequence 20 to 998. The torsion angle ϕ is not defined for the first location and ψ for the last location. Excluding the first and last locations and other missing values, the SADS-CoV data set we analyze is the sequence of $(\phi_i, \psi_i) \in \mathbb{T}^2$ of length $n = 964$. Similar preprocessing gives HKU2 and SARS-CoV-2 torsion angles of sizes 964 and 824. Note that each of these three torsion angles data sets is sequentially observed along the backbone of a protein. The torsion angles are thus, in fact, serially correlated, and the assumption of exchangeability may not hold in general. Nevertheless, inspecting the scatterplot of the (ϕ_i, ψ_i) angles has been useful in understanding a protein structure (Eltzner, Huckemann and Mardia (2018), Mardia, Taylor and Subramaniam (2007)). Our analysis only uses the scatter of torsion angles which can then be assumed independent.

Prediction sets and the choice of tuning parameters. Since the sample size $n = 964$ is not small, it makes sense to construct the inductive conformal prediction sets. We demonstrate the use of conformity scores based on kernel density estimates and the estimated bivariate von Mises mixtures. The concentration parameter $\kappa > 0$ plays the role of the reciprocal of the bandwidth in the usual kernel density estimation. For large κ, the prediction set $\hat{C}_n^{\mathrm{kde}}(\alpha, \kappa)$ is prone to catch random fluctuations in the data set, as exemplified for $\hat{C}_n^{\mathrm{kde}}(0.1, 100)$, shown in the bottom right panel of Figure 5. The difference of $\hat{C}_n^{\mathrm{kde}}(\alpha, \kappa)$, according to varying κ, is less severe if the coverage $1 - \alpha$ is high. As an instance, the 98% prediction sets shown in the

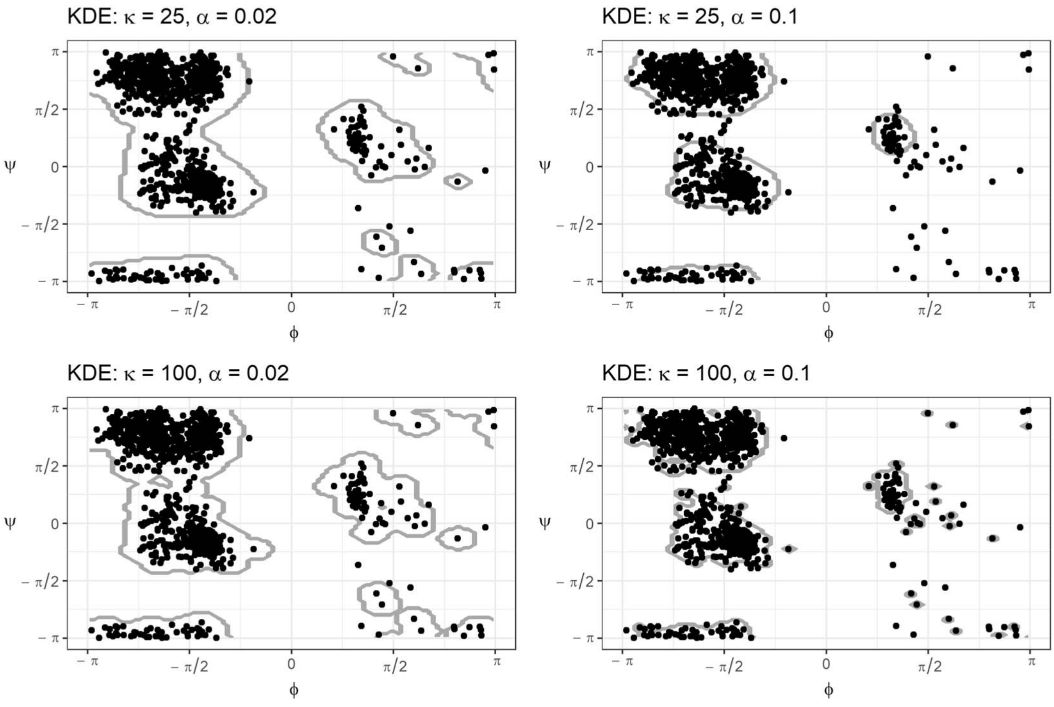


FIG. 5. *Conformal prediction sets $\hat{C}_n^{\mathrm{kde}}(\alpha, \kappa)$, the boundaries of which are displayed as gray curves, for the SADS-CoV torsion angles.*

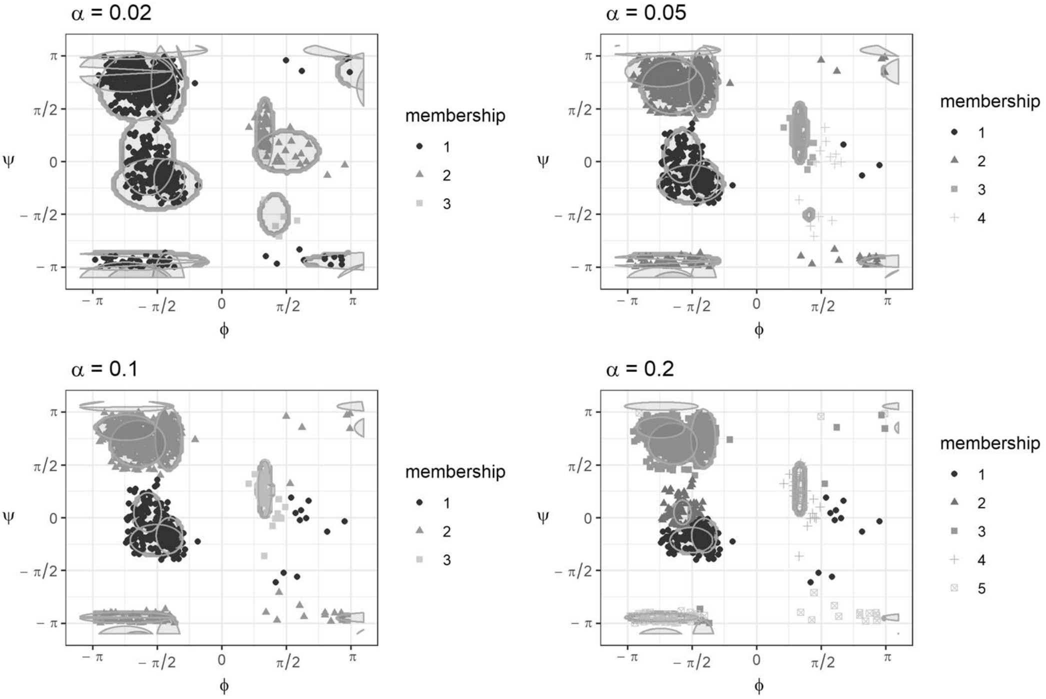


FIG. 6. *Conformal prediction sets $\hat{C}_n^e(\alpha, 10)$, their elliptical components and cluster memberships for torsion angles of the SADS-CoV protein structure.*

left panels of Figure 5 are similar to each other, in spite of different choices of $\kappa = 25, 100$. Choosing $(\alpha, \kappa) = (0.1, 25)$ gives a nice separation of three connected components of the prediction sets (top right panel).

A bivariate von Mises mixture distribution with $J = 10$ components is also fitted to the data with the restriction of $\lambda = 0$. (Other choices of J are considered later.) Prediction sets are obtained from this mixture model at various nominal level $1 - \alpha$. For cluster identification and assignment, the prediction set $\hat{C}_n^e$, consisting of the union of ellipses, is convenient in identifying the connected components. To inspect the effect of varying α, we have plotted $\hat{C}_n^e = \bigcup_{j=1}^J \hat{E}_j(\sigma_{(i_{n_2},\alpha)})$ and each of the ellipses ($\hat{E}_j$'s) for a few choices of α in Figure 6. When the coverage $1-\alpha$ is large, the ellipses involved are large as well, resulting in a smaller number of connected components. As α increases, the radii of all ellipses decrease which results in a connected component either to be divided into two components or to disappear. For example, the largest connected component (labeled 1) for the $\alpha = 0.02$ case is divided into two connected components (labeled 1 and 2) for the $\alpha = 0.05$ case; see the top two panels of Figure 6. On the other hand, comparing $\alpha = 0.05$ with 0.1 (bottom left panel), the connected component labeled 4 at $\alpha = 0.05$ disappears at $\alpha = 0.1$, as the ellipse becomes an empty set. For each choice of α, the number K of clusters are counted as the number of connected components, and the cluster membership $A_e(x)$ (11) is computed for all torsion angles x in the SADS-CoV protein structure, also shown in Figure 6. It is evident that the clustering results depend on the choice of α and the hyperparameter J used in fitting the mixture model.

If there is a desired level of coverage $1 - \alpha$, then the best hyperparameter J (or κ) can be chosen to minimize the volume $\mu(C)$ of the corresponding prediction set C (13). For the SADS-CoV torsion angles, the graph of $(\kappa, \mu(C))$ is roughly convex (as shown in the top left panel of Figure 7), leading that a choice of κ near 30 is a stable choice. Note that the kde-based prediction set $\hat{C}_n^{\text{kde}}(\alpha, \kappa)$ is continuous with respect to both α and κ. On the other

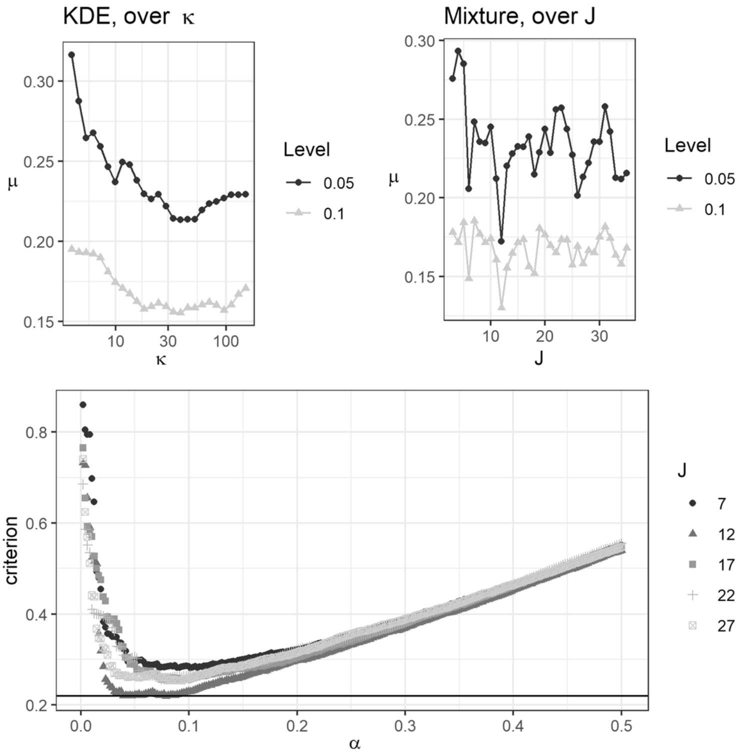


FIG. 7. *The choice of level α and the hyperparameter J or κ is given by the minimizer of the criterion* (15), *and is shown for SADS-CoV torsion angles. The bottom panel shows the values of criterion for a few choices of J.* (*This figure with all values of J is contained in the Appendix Section C in the Supplementary Material* (*Jung, Park and Kim* (2021)).)

hand, the volume of the mixture-model based prediction set $\mu\{\hat{C}_n^e(\alpha, J)\}$ has a large fluctuation over increasing J (see the top-right panel of Figure 7), indicating that estimating the mixture model with a large number of components from a sample of size $n_1 = 482$ might lead to unstable estimates. For the SARS-CoV-2 torsion angles, the graph of $(\alpha, \mu\{\hat{C}_n^e(\alpha, J)\})$ is roughly convex and smooth (see Appendix Figure C.4 in the Supplementary Material (Jung, Park and Kim (2021))). Nevertheless, in view of the criterion $\nu(\alpha, J) := \alpha + \mu(\hat{C}_n)$ (15), changes due to the level is more substantial than the changes over J. This is illustrated in the bottom panel of Figure 7, where the values of $\nu(\alpha, J)$ are evaluated and plotted for various combinations of $\alpha \in (0, 0.5)$ and a few choices of $J \in \{3, \dots, 35\}$. The criterion was minimized to be $\nu(\hat{\alpha}, \hat{J}) \approx 0.241$ at $\hat{\alpha} \approx 0.079$, $\hat{J} = 12$, at which more than 92% of the sample is bound to lie in the prediction set $\hat{C}_n$ whose volume is $100\mu(\hat{C}_n)\% \approx 14.1\%$ of the torus. Here, $\hat{C}_n = \hat{C}_n^e$. The level $\hat{\alpha}$ and the number of components $\hat{J} = 12$ are used in the subsequent analysis of clustering for SADS-CoV torsion angles. A similar analysis done for SARS-CoV-2 leads $\hat{\alpha} \approx 0.10$, $\hat{J} = 22$.

Clustering and a post analysis. For each set of SADS-CoV and SARS-CoV-2 torsion angles, we obtained $K = 3$ clusters $\mathcal{E}_k$ for the SADS-CoV ($K = 6$ clusters for the SARS-CoV-2) and evaluated the cluster membership $A_e(x)$ (11) for each data set; see Figure 8.

The protein structure is known to follow strict geometric rules. The three larger clusters (shown in both cases) correspond to the well-known shapes of protein structures (Lovell et al. (2003), Walther and Cohen (1999)). In particular, The cluster labeled 1 in the top panel of Figure 8 is related to the right-handed α-helix, Cluster 2 is to the β-sheet and Cluster 3 is to the left-handed α-helix. In the bottom panel of Figure 8 for the SARS-CoV-2, three

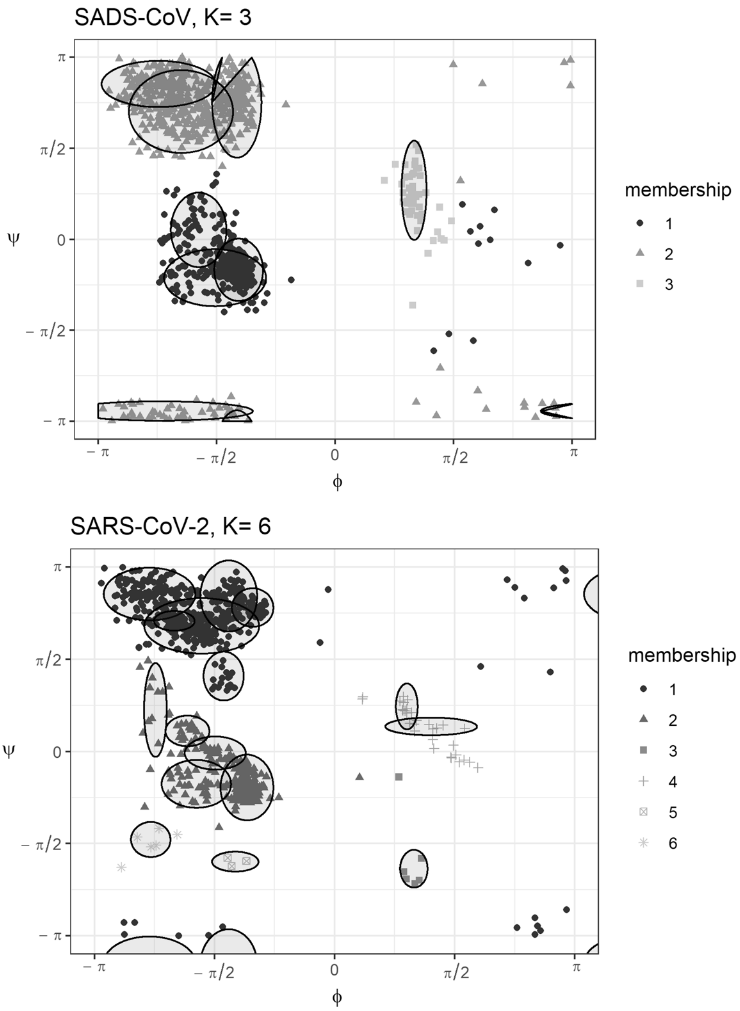


FIG. 8. *Clustering results for SADS-CoV torsion angles (top) and SARS-CoV-2 torsion angles (bottom).*

additional smaller clusters are shown. It appears that Cluster 3 (filled squares) corresponds to a less common conformation called γ-turns (cf. Figure 1 of Lovell et al. (2003)). Clusters 5 and 6 of the SARS-CoV-2 do not overlap with any torsion angles of SADS-CoV, indicating a potential difference in protein structures of the two coronaviruses.

Empirical coverage of prediction regions. The SADS-CoV protein structure is known to be nearly identical to the HKU2 (Yu et al. (2020)). On the other hand, SARS-CoV-2 is a different strain of coronavirus, whose protein backbone structure differs from the others. The prediction sets $\hat{C}_n$, based on which the proposed clustering is conducted, are theoretically valid, that is, $P(x \in \hat{C}_n) \geq 1 - \alpha$, for all $\alpha \geq 0$, for any n and for any choice of prediction set $\hat{C}_n^{\text{kde}}$, $\hat{C}_n^{e}$, $\hat{C}_n^{\text{mix}}$ and $\hat{C}_n^{\text{max}}$. We empirically confirm that the coverage of $\hat{C}_n$ (estimated from the SADS-CoV angles) meets the nominal level $1 - \alpha$ for the SADS-CoV and HKU2 torsion angles. In the top two panels of Figure 9, the empirical coverage probability $\hat{P}(x \in \hat{C}_n)$ is near the nominal level $1 - \alpha$ and is generally above the lower bound of a 98% pointwise confidence interval for $P(x \in \hat{C}_n \mid \hat{C}_n)$, for most values of $\alpha \in (0, 0.2)$.

For the SARS-CoV-2 angles, shown in the bottom panel of Figure 9, the coverage of the SADS-CoV-based prediction sets is below the 98% pointwise confidence lower bound for higher levels $1 - \alpha > 0.9$. This indicates again a potentially different backbone structure of

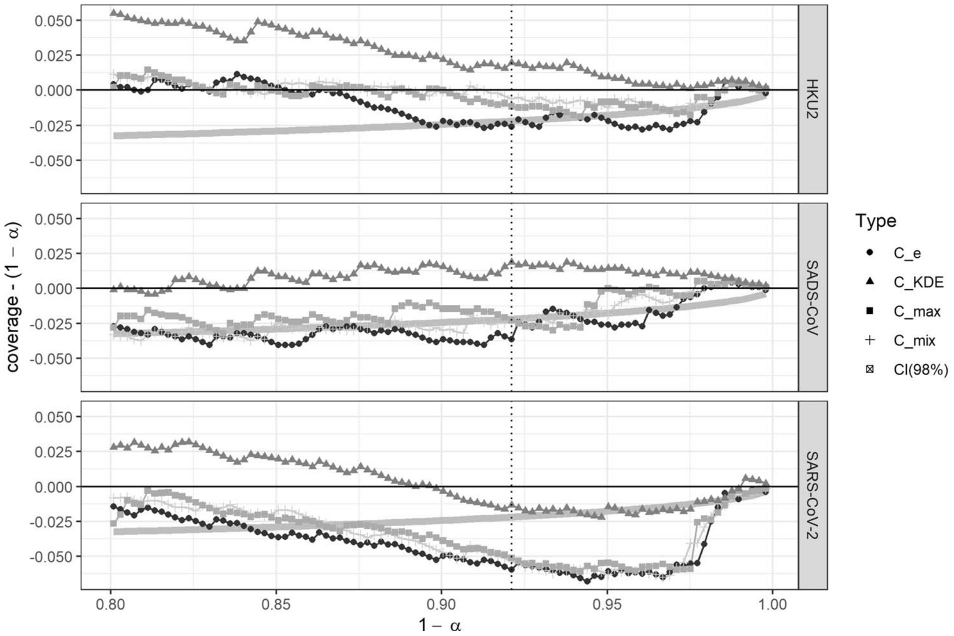


FIG. 9. *Empirical coverage of the prediction sets* (*estimated from the SADS-CoV torsion angles*) *for the HKU*2, *SADS-CoV and SARS-CoV*-2 *torsion angles* (*shown on top*, *middle and bottom panels*, *respectively*). *The vertical dotted line is at* $1-\hat{\alpha}$, *at which the clustering for SADS-CoV angles is performed with* $\hat{C}_n^e(\hat{\alpha}, \hat{J})$.

the SARS-CoV-2 protein. Caution is needed in making a conclusion since the confidence limits are obtained for the given prediction set. In Appendix B.4 in the Supplementary Material (Jung, Park and Kim (2021)), we show that the variance due to the estimation of $\hat{C}_n$ is larger than the variance due to the evaluation of the coverage, from a simulated data set of Section 4.2.

The hyperparameters for the prediction sets in Figure 9 are chosen to be $\hat{\kappa} = 30$ and $\hat{J} = 12$ based on the criterion (15). Other choices of κ and J lead to similar results.

4.2. *Clustering artificial toroidal data.* The proposed clustering procedure is tested upon two artificial data on the torus. We empirically compare the clustering performances of our proposal with existing methods in Section 3.1.

The two artificial data are each sampled from the following models:

- Model I: The dataset of size $n = 270$ is sampled from a mixture of $K = 5$ clusters, where three clusters are sampled from bivariate normal distributions (with sizes 70, 50, 50), and the other two are each sampled from the uniform distribution on a rectangle defined on $\mathbb{R}^2$ (each with size 50), then wrapped onto the torus.
- Model II: The dataset of size $n = 500$ is sampled from a mixture of $K = 3$ clusters, where the first cluster is sampled from a spherical normal distribution with size $n_1 = 100$, the second cluster of size $n_2 = 350$ is from the uniform distribution on a large “L”-shaped region and the third cluster of size 50 is sampled from the uniform distribution on the entire $\mathbb{T}^2$.

Data sets described above are generated for the estimation of prediction regions and clustering rules (displayed in Figure 10). These data sets are called training data. Independent sets of data from the same models are also generated for validation and are called testing data. Clustering rules based on the mixture models with J, α chosen by the proposed criterion (15) result in $K = 5$ clusters for Model I and $K = 2$ for Model II.

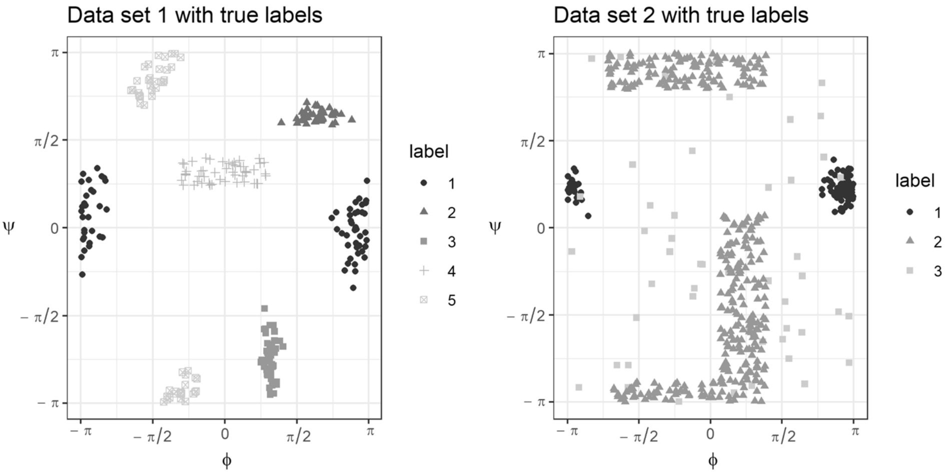


FIG. 10. *Simulated toroidal data sets with true cluster labels.*

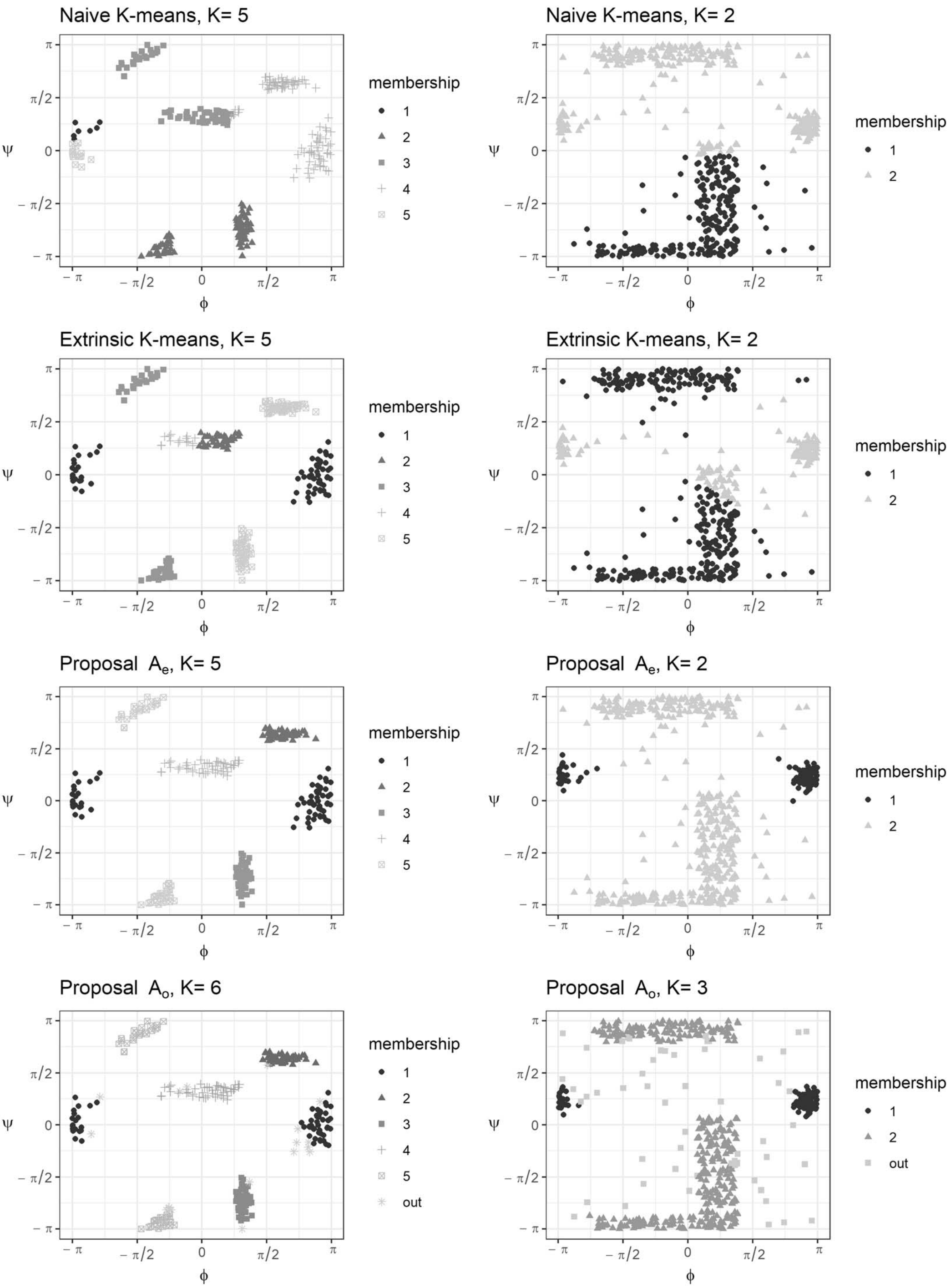
We have compared four methods of clustering: (*i*) the naive k-means clustering (ignoring the angular constraint), (*ii*) the extrinsic k-means clustering in the ambient space and the proposed predictive clustering with membership assignment by (*iii*) the clustering rule A_e (11) and by (*iv*) the clustering rule A_o (12), creating an "outlier" cluster; see Sections 3.1 and 3.2 for details. For each model the clustering rules are estimated using the training data, based on which the cluster memberships of the testing data are evaluated. Figure 11 displays the predicted cluster memberships for the testing data.

We now inspect the results for each model. Since Model I consists of five well-separated clusters, human eye is excellent in accurately clustering the data with an understanding of the "cut-and-flattened" torus plots shown in Figures 10–11. The naive k-means algorithm does not reflect the geometry of toroidal data and results in a poor clustering (top left panel of Figure 11). The extrinsic k-means correctly identifies the cluster #1 (split across $\phi = -\pi = \pi$) but mistakenly splits the elongated cluster (labeled 4 in Figure 10) into two clusters (labeled 2 and 4 in Figure 11). Our proposal correctly predicts all cluster memberships. To quantify the quality of clustering, we use the adjusted Rand index (Hubert and Arabie (1985)). For comparison of two cluster indices, one with G clusters and the other with K clusters, let N_{gk} be the number of observations whose first cluster index is g and the second index is k. The index is defined as

$$\mathrm{ARI} = \frac{\sum_{g=1}^{G} \sum_{k=1}^{K} \binom{N_{gk}}{2} - N_G N_K / \binom{N}{2}}{(N_G + N_K)/2 - N_G N_G / \binom{N}{2}},$$

where $N_G = \sum_{g=1}^{G} \binom{N_{g\cdot}}{2}$ and $N_K = \sum_{k=1}^{K} \binom{N_{\cdot k}}{2}$. The ARI has the value of 1 when two indices match perfectly. Table 1 collects the ARIs comparing the predicted labels from each clustering method to the true cluster labels of the testing data. The proposed clustering with assignment rule A_e performs almost perfectly for data from Model I.

The data from Model II can be viewed as two large clusters sprinkled with outliers (see right panels of Figure 11). For this reason we have used $K = 2$ in the k-means clusterings. As before, the naive k-means algorithm is not suitable for toroidal data. The extrinsic k-means is not successful for this data set due to the irregularly shaped cluster. The proposed clustering methods correctly find the two large clusters. When the membership is assigned by A_e (using the conformity score $\hat{e}_j(x)$), all observations belonging to the true "outlier" group are forced to be assigned to either the "L"-shaped cluster or the ball-shaped cluster. On the other hand, the cluster membership rule A_o with outliers performs better, as evidenced by the highest ARI in Table 1 and the visually satisfying clustering result in the lower right panel of Figure 11.


Naive K-means, K= 5

Naive K-means, K= 2

Extrinsic K-means, K= 5

Extrinsic K-means, K= 2

Proposal A_e, K= 5

Proposal A_e, K= 2

Proposal A_o, K= 6

Proposal A_o, K= 3

FIG. 11. *Clustering results for testing sets. Results for Model I are shown on the left column, Model II on the right.*

5. Summary and discussion. We have introduced a clustering procedure for toroidal data based on the conformal prediction sets. Conformal prediction sets can be estimated using any conformity score, and we have demonstrated the use of the kernel density estimates, density estimates from a bivariate von Mises mixture model and their variants in the construction of the conformity scores. The proposed clustering uses the mixture model-based

TABLE 1

Adjusted Rand indices (ARI), evaluated for predicted cluster labels of synthetic toroidal data. Shown are the averages (and standard deviations in the parenthesis) of the ARIs obtained from 100 *repetitions. The higher ARI, the better clustering membership prediction*

Model	Naive K-means	Extrinsic K-means	Proposal (A_e)	Proposal (A_o)
I	0.57 (0.06)	0.89 (0.13)	0.98 (0.03)	0.87 (0.06)
II	0.12 (0.10)	0.43 (0.13)	0.73 (0.07)	0.82 (0.09)

prediction sets in identifying clusters and cluster membership assignment. We have shown that our approach performs better than naive adaptations of off-the-shelf clustering algorithms to toroidal data.

Our approach can be easily adapted for clustering the usual multivariate data in low-dimensional vector spaces. For data in the Euclidean space, the idea of using conformal prediction framework for clustering has also been explored by Shin, Rinaldo and Wasserman (2019) and Nouretdinov et al. (2020). Our proposal shares some similarity with these works but differs in the choices of conformal scores and the hyperparameters (as well as the different sample spaces). We also point out that the goal of Shin, Rinaldo and Wasserman (2019), building a prediction set based on clustering, is the opposite of this work: Cluster assignment based on prediction sets.

Our approach can be naturally extended to handle data on general tori $\mathbb{T}^p$ ($p \geq 2$), directional data on hyperspheres $\{x \in \mathbb{R}^p : \|x\|_2 = 1\}$ or other types of manifold-valued data. We have focused on the bivariate circular variables $(\phi, \psi) \in \mathbb{T}^2$, since the torsion angles have been known to well represent major backbone structures of proteins. For multivariate angles in a higher dimensional torus, fitting mixture models can be computationally expensive and unstable. A key direction for future research is to devise a conformity score that is computationally efficient for high p, and is easy to identify the resulting connected components (clusters). General tori $\mathbb{T}^p$ appear as the data space for some biochemical applications. For example, the backbone of a protein has, in fact, a third angle, called ω-angle, which are often ignored because $\omega \approx \pi$. If the protein has side chains, at each side chain there are a few torsion angles χ_i defined. In describing richer RNA structures, Sargsyan, Wright and Lim (2012) and Eltzner, Huckemann and Mardia (2018) have used seven torsion angles of RNA structures in $\mathbb{T}^7$.

An important issue we have not fully investigated is the sampling distribution of $\hat{C}_n$. An empirical investigation in Appendix B.4 in the Supplementary Material (Jung, Park and Kim (2021)) indicates that the variation due to $\hat{C}_n$ in $P(x \in \hat{C}_n \mid \hat{C}_n)$ is substantial. We leave this as a future topic of research.

Acknowledgments. We thank the Editor, the Associate Editor and two anonymous referees for their constructive comments which helped us to substantially improve our manuscript.

Funding. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C2002256).

SUPPLEMENTARY MATERIAL

Supplement to "Clustering on the torus by conformal prediction" (DOI: 10.1214/21-AOAS1459SUPPA; .pdf). We provide (A) EM algorithms for mixtures of bivariate von Mises and an algorithm to test the intersection of two toroidal ellipses, (B) proofs and technical details, and (C) supplementary figures referenced in Section 4.

Source code and data for "Clustering on the torus by conformal prediction" (DOI: 10.1214/21-AOAS1459SUPPB; .zip). Data and R codes to reproduce the analysis are contained in a zipped file RcodesClustorus.zip. R functions used in the analysis are also available at https://github.com/sungkyujung/ClusTorus.

REFERENCES

ARTHUR, D. and VASSILVITSKII, S. (2007). `k-means++`: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 1027–1035. ACM, New York. MR2485254

BERG, J. M., TYMOCZKO, J. L. and STRYER, L. (2002). *Biochemistry*, 5th ed. W. H. Freeman & Company, New York.

BLUM, H. (1967). A transformation for extracting new descriptors of shape. In *Models for the Perception of Speech and Visual Form* (W. Wathen-Dunn, ed.) 362–380. MIT Press, Cambridge.

CHAKRABORTY, S. and WONG, S. W. (2017). BAMBI: An R package for fitting bivariate angular mixture models. arXiv preprint arXiv:1708.07804.

CHAN, J. F.-W., YUAN, S., KOK, K.-H., TO, K. K.-W., CHU, H., YANG, J., XING, F., LIU, J., YIP, C. C.-Y. et al. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster. *Lancet* **395** 514–523.

CHENG, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **17** 790–799.

DI MARZIO, M., PANZERA, A. and TAYLOR, C. C. (2011). Kernel density estimation on the torus. *J. Statist. Plann. Inference* **141** 2156–2173. MR2772221 https://doi.org/10.1016/j.jspi.2011.01.002

DILL, K. A. and MACCALLUM, J. L. (2012). The protein-folding problem, 50 years on. *Science* **338** 1042–1046.

ELTZNER, B., HUCKEMANN, S. and MARDIA, K. V. (2018). Torus principal component analysis with applications to RNA structure. *Ann. Appl. Stat.* **12** 1332–1359. MR3834306 https://doi.org/10.1214/17-AOAS1115

GAO, Y., WANG, S., DENG, M. and XU, J. (2018). RaptorX-Angle: Real-value prediction of protein backbone dihedral angles through a hybrid method of clustering and deep learning. *BMC Bioinform.* **19** 100.

GONG, L., LI, J., ZHOU, Q., XU, Z., CHEN, L., ZHANG, Y., XUE, C., WEN, Z. and CAO, Y. (2017). A new bat-HKU2-like coronavirus in swine, China, 2017. *Emerg. Infec. Dis.* **23** 1607.

GORBALENYA, A. E., BAKER, S. C., BARIC, R. S. and CORONAVIRIDAE STUDY GROUP OF THE INTERNATIONAL COMMITTEE ON TAXONOMY OF VIRUSES (2020). The species severe acute respiratory syndrome-related coronavirus: Classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* **5** 536.

GRANT, B. J., RODRIGUES, A. P., ELSAWY, K. M., MCCAMMON, J. A. and CAVES, L. S. (2006). Bio3d: An R package for the comparative analysis of protein structures. *Bioinformatics* **22** 2695–2696.

HARTIGAN, J. A. (1975). *Clustering Algorithms. Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0405726

HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.

JUNG, S., PARK, K. and KIM, B. (2021). Supplement to "Clustering on the torus by conformal prediction." https://doi.org/10.1214/21-AOAS1459SUPPA, https://doi.org/10.1214/21-AOAS1459SUPPB

KAUFMAN, L. and ROUSSEEUW, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis* **344**. John Wiley & Sons. MR1044997 https://doi.org/10.1002/9780470316801

KOUNTOURIS, P. and HIRST, J. D. (2009). Prediction of backbone dihedral angles and protein secondary structure using support vector machines. *BMC Bioinform.* **10** 437. https://doi.org/10.1186/1471-2105-10-437

LEI, J., RINALDO, A. and WASSERMAN, L. (2015). A conformal prediction approach to explore functional data. *Ann. Math. Artif. Intell.* **74** 29–43. MR3353895 https://doi.org/10.1007/s10472-013-9366-6

LEI, J., ROBINS, J. and WASSERMAN, L. (2013). Distribution-free prediction sets. *J. Amer. Statist. Assoc.* **108** 278–287. MR3174619 https://doi.org/10.1080/01621459.2012.751873

LEI, J., G'SELL, M., RINALDO, A., TIBSHIRANI, R. J. and WASSERMAN, L. (2018). Distribution-free predictive inference for regression. *J. Amer. Statist. Assoc.* **113** 1094–1111. MR3862342 https://doi.org/10.1080/01621459.2017.1307116

LENNOX, K. P., DAHL, D. B., VANNUCCI, M. and TSAI, J. W. (2009). Density estimation for protein conformation angles using a bivariate von Mises distribution and Bayesian nonparametrics. *J. Amer. Statist. Assoc.* **104** 586–596. MR2751440 https://doi.org/10.1198/jasa.2009.0024

LOVELL, S. C., DAVIS, I. W., ARENDALL III, W. B., DE BAKKER, P. I., WORD, J. M., PRISANT, M. G., RICHARDSON, J. S. and RICHARDSON, D. C. (2003). Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. *Proteins: Structure, Function, and Bioinformatics* **50** 437–450.

MARDIA, K. V. and JUPP, P. E. (2000). *Directional Statistics. Wiley Series in Probability and Statistics*. Wiley, Chichester. Revised reprint of *Statistics of directional data* by Mardia [MR0336854 (49 #1627)]. MR1828667

MARDIA, K. V., TAYLOR, C. C. and SUBRAMANIAM, G. K. (2007). Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics* **63** 505–512. MR2370809 https://doi.org/10.1111/j.1541-0420.2006.00682.x

MARDIA, K. V., HUGHES, G., TAYLOR, C. C. and SINGH, H. (2008). A multivariate von Mises distribution with applications to bioinformatics. *Canad. J. Statist.* **36** 99–109. MR2432195 https://doi.org/10.1002/cjs.5550360110

MARDIA, K. V., KENT, J. T., ZHANG, Z., TAYLOR, C. C. and HAMELRYCK, T. (2012). Mixtures of concentrated multivariate sine distributions with applications to bioinformatics. *J. Appl. Stat.* **39** 2475–2492. MR2993298 https://doi.org/10.1080/02664763.2012.719221

MURTAGH, F. and CONTRERAS, P. (2012). Algorithms for hierarchical clustering: An overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2** 86–97.

MURTAGH, F. and CONTRERAS, P. (2017). Algorithms for hierarchical clustering: An overview, II. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **7** e1219.

NODEHI, A., GOLALIZADEH, M., MAADOOLIAT, M. and AGOSTINELLI, C. (2021). Estimation of parameters in multivariate wrapped models for data on a p-torus. *Comput. Statist.* **36** 193–215. MR4215388 https://doi.org/10.1007/s00180-020-01006-x

NOURETDINOV, I., GAMMERMAN, J., FONTANA, M. and REHAL, D. (2020). Multi-level conformal clustering: A distribution-free technique for clustering and anomaly detection. *Neurocomputing* **397** 279–291.

O'NEILL, B. (2006). *Elementary Differential Geometry*, 2nd ed. Elsevier/Academic Press, Amsterdam. MR2351345

POLONIK, W. (1997). Minimum volume sets and generalized quantile processes. *Stochastic Process. Appl.* **69** 1–24. MR1464172 https://doi.org/10.1016/S0304-4149(97)00028-8

SARGSYAN, K., WRIGHT, J. and LIM, C. (2012). GeoPCA: A new tool for multivariate analysis of dihedral angles based on principal component geodesics. *Nucleic Acids Res.* **40** e25–e25.

SCRUCCA, L., FOP, M., MURPHY, T. B. and RAFTERY, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **8** 289–317.

SHAPOVALOV, M., VUCETIC, S. and DUNBRACK JR., R. L. (2019). A new clustering and nomenclature for beta turns derived from high-resolution protein structures. *PLoS Comput. Biol.* **15** e1006844.

SHIN, J., RINALDO, A. and WASSERMAN, L. (2019). Predictive clustering. arXiv preprint arXiv:1903.08125.

SINGH, H., HNIZDO, V. and DEMCHUK, E. (2002). Probabilistic model for two dependent circular variables. *Biometrika* **89** 719–723. MR1929175 https://doi.org/10.1093/biomet/89.3.719

VAN DER LAAN, M. J., POLLARD, K. S. and BRYAN, J. (2003). A new partitioning around medoids algorithm. *J. Stat. Comput. Simul.* **73** 575–584. MR1998670 https://doi.org/10.1080/0094965031000136012

VOVK, V., GAMMERMAN, A. and SHAFER, G. (2005). *Algorithmic Learning in a Random World*. Springer, New York. MR2161220

WALLS, A. C., PARK, Y.-J., TORTORICI, M. A., WALL, A., MCGUIRE, A. T. and VEESLER, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*.

WALTHER, D. and COHEN, F. E. (1999). Conformational attractors on the Ramachandran map. *Acta Crystallogr., Sect. D, Biol. Crystallogr.* **55** 506–517.

XU, D. and TIAN, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science* **2** 165–193.

YU, J., QIAO, S., GUO, R. and WANG, X. (2020). Cryo-EM structures of HKU2 and SADS-CoV spike glycoproteins provide insights into coronavirus evolution. *Nat. Commun.* **11** 1–12.