International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol.4 Issue 9, September-2020, Pages: 18-25

Kernel Density Estimation on the Torus with Application to Bioinformatics

Samira Faisal Abushilah & Hayder Ali Hussein

Department of Mathematics, Faculty of Education for Girls, University of Kufa, Najaf, Iraq.
Email: Samirafaisal80@gmail.com & haideralitap2@gmail.com

Abstract: *In this paper, we suggest an approach to estimate the probability density function for bivariate circular data. Kernel density estimation technique with different concentration parameters and different kernel functions are used to construct this approach. In our approach, we used Wrapped Cauchy distribution and Wrapped Normal distribution as a kernel function. Moreover, the procedure that we have suggested is applied to two types of data (simulated data and protein data). Firstly, we apply the approach that we have suggested to simulated data, which are generated from the bivariate von-Mises distribution, to highlight the joint probability density function for the bivariate von-Mises distribution. Secondly, the proposed approach is applied to real data, from the Protein Data Bank (PDB), to show the joint probability density function for some types of proteins.*

Keywords: Bivariate Circular Data; Kernel Density Estimation; Protein Dihedral Angles; Wrapped Cauchy distribution; Wrapped Normal distribution.

1. Introduction

Circular data can be represented by an angle $\boldsymbol{\theta}$ in [0, $2\boldsymbol{\pi}$) and can be seen as a point on the unit circle. This data is periodic (i.e. $\boldsymbol{\theta = \theta + 2n\pi}$, where $\boldsymbol{n \in Z}$) and this type of data differs from traditional linear data and statistical methods to handle this type of data are relatively new and are still under development [6]. Circular data occur frequently and in many cases we need to know the joint distribution of two or more circular random variables. For example, over a time period and in the study of wind directions we look for a model to describe a bivariate circular data. In fact, with simple transformation the temporal variables are converted into circular variables [3].

On a data lying on the circle or on the sphere there are a few contributions. In 1987, Hall et al. have been used directional derivatives to study kernel density estimators for spherical data [5]. Fisher (1989) was the first to propose a kernel estimator by adapting the method of Silverman (1986) for linear data [8] and they have used a quartic kernel function K($\boldsymbol{\theta}$)=0.9375(1−$\boldsymbol{\theta}$^2)^2 [4] (note this is not a circular kernel). However, for kernel density estimation on the torus nothing specifically until 2011.

Di Marzio et al. (2011) extended the circular kernel density estimator to a p-dimensional torus and they considered cross-validation and bootstrap approach for bandwidth selection under the assumption that the concentration parameters for the kernel functions are the same ($\boldsymbol{\kappa_1 = \kappa_2}$) [3]. Under the same assumption, Taylor et al. (2012) proposed a procedure to evaluate a protein structure by computing a tail probability which is based on the amino acid kernel density estimates [10]. In fact, estimation of the kernel density on the torus needs to determine bivariate kernel function with different concentration parameters (smoothing parameters) to get an accurate estimation because the optimal bandwidth of kernel density estimation depends on the second density derivative [9]. Since the second derivative in general will be different in each direction, then we expect the optimal pair of smoothing parameters to be unequal. Therefore, we need to develop an approach to estimate the kernel density function for bivariate circular data with different concentration parameters for the kernel function.

In recent years, there are a few contributions that focus on fitting von Mises distribution and wrapped normal distribution for circular data. In 2014, Hornik et al. wrote an R package, movMF, containing functionality for fitting finite mixtures of von Mises-Fisher distributions, using the expectation maximisation algorithm (EM) for maximum likelihood estimation [6]. Chakraborty et al. (2017) introduced an R package, BAMBI. This package provides Bayesian methods to model univariate and bivariate angular data for some circular distributions such as von Mises and wrapped normal using finite mixture models of these distributions [2]. Abushilah (2019) proposed a methodology to estimate probability density function for bivariate circular data using von Mises distribution as a kernel function, and they applied this methodology to highlight the probability density function for 20 amino acids from PDB [1]. Although the above two packages are good steps for fitting bivariate von Mises and normal wrapped distribution, we seek to estimate the probability density function for bivariate circular data on the torus without making assumptions about the distribution of the data.

Therefore, we suggest in this paper an approach to estimate the probability density function for bivariate circular data ($\boldsymbol{\phi_i, \psi_i}$), i=1, 2, ..., m, using kernel density estimation with different concentration parameters. This probability density function could be used to highlight the joint probability density function for dihedral angles which belong to some proteins.

We begin in Section 2 by estimating the joint probability density function for bivariate circular data. In Section 3, we will apply the approach that we have proposed in Section 2 to simulated data and to protein data to highlight the probability density function for the dihedral angles $(\boldsymbol{\phi}, \boldsymbol{\psi})$ which belong to some proteins. Finally, the conclusion of this paper will be presented.

2. Kernel Density on the Torus

Kernel density estimation is a non-parametric method to estimate the probability density function of a random variable. This method relies on a mapping between two spaces: input space and feature space, where the kernel function is used to compute the inner product of the vectors in the feature space [11].

In this paper, we suggest an approach to estimate the probability density function for bivariate circular data $(\boldsymbol{\phi_i}, \boldsymbol{\psi_i})$, i=1, 2, ..., m. Kernel density estimation technique with different concentration parameters and different kernel functions are used to construct this approach. In our approach, we used Wrapped Cauchy distribution and Wrapped Normal distribution as a kernel function.

The wrapped Cauchy distribution WC$(\boldsymbol{\mu}, \boldsymbol{\mathcal{E}})$ is a symmetric unimodal distribution with the probability density function which is given by

$$\boldsymbol{f(\theta;\ \mu, \mathcal{E}) = \frac{1}{2\pi}\ \frac{1-\mathcal{E}^2}{1+\mathcal{E}^2-2\mathcal{E}\cos(\theta-\mu)}, -\pi \le \theta \le \pi, 0 \le \mathcal{E} \le 1,} \qquad (1)$$

where the parameter $\boldsymbol{\mu}$ represents the mean direction and $\boldsymbol{\mathcal{E}}$ represents concentration parameter of the distribution.

The wrapped Normal distribution WN$(\boldsymbol{\mu}, \boldsymbol{\mathcal{E}})$ is a symmetric unimodal two-parameter distribution with probability density function which is given by

$$\boldsymbol{f(\theta;\ \mu, \mathcal{E}) = \frac{1}{2\pi}\left(1+2\sum_{p=1}^{\infty}\mathcal{E}^{p^2}\cos p(\theta-\mu)\right), -\pi \le \theta \le \pi, 0 \le \mathcal{E} \le 1,} \qquad (2)$$

where $\boldsymbol{\mu}$ represents the mean direction and $\boldsymbol{\mathcal{E}}$ represents the mean resultant length of the distribution.

For bivariate circular data $(\boldsymbol{\phi_1},\ \boldsymbol{\psi_1}), (\boldsymbol{\phi_2}, \boldsymbol{\psi_2}), \ldots, (\boldsymbol{\phi_m}, \boldsymbol{\psi_m})$, a kernel density estimate $\boldsymbol{\hat{f}(\phi, \psi)}$ can be constructed by fixing a circular density for the kernel with mean zero and concentration parameter $\boldsymbol{\mathcal{E}}$ corresponding to the smoothing parameter, the kernel density estimation is given by

$$\boldsymbol{\hat{f}(\phi,\psi) = \frac{1}{m}\sum_{i=1}^{m} K(\phi-\phi_i;\ \mu, \mathcal{E}_1) K(\psi-\psi_i;\ \mu, \mathcal{E}_2),} \qquad (3)$$

where $(\boldsymbol{\mathcal{E}_1},\ \boldsymbol{\mathcal{E}_2})$ are the smoothing parameters (bandwidth) and $\boldsymbol{K}(\cdot)$ is a kernel function, which is given in Equations (1) and (2).

As we can see in the right hand side of Equation (3) there is a multivariate kernel. This multivariate kernel is used for the reason that the data is bivariate, and there is a specific kernel with different parameters for each of the observation. Using a multiplicative kernel is not equivalent to a summing independence of the variables mentioned by Silverman (1986).

After the above methodology has been constructed, it is applied to protein data in order to highlight the probability density function for the dihedral angles $(\boldsymbol{\phi}, \boldsymbol{\psi})$ which belong to some proteins. We show the proposed approach in Algorithm 1.

Algorithm 1: Kernel density estimatin for the bivariate circular data with Wrapped Cauchy kernel.

Data: Bivariate circular data $\{(\phi_1, \psi_1),\ (\phi_2, \psi_2),\ .....,\ (\phi_m, \psi_m)\}$.
Result: The joint probability density function for the data.
Function `f_hat`$(phi,\ psi,\ h1,\ h2,\ n_1,\ lims = c(range(x),\ range(y)))$**:**
  m=length(phi)
  ph=sequence(lims[1], lims[2], length.out=n_1)
  ps=sequence(lims[3], lims[4], length.out= n_1)
  calculate $\triangle x = ph - phi$
  calculate $\triangle y = ps - psi$
  for each combination of $\phi = \{\phi_1,\ \phi_2,\ .....,\ \phi_m\}$ sequence and $\psi = \{\psi_1,\ \psi_2,\ .....,\ \psi_m\}$ sequence calculate $\widehat{f}$ using $\widehat{f}(\phi,\ \psi) = \frac{1}{m}\sum_{i=1}^{m}[K(\triangle x;\ 0,\ h_1)K(\triangle y;\ 0,\ h_2)]$, where K is Wrapped Cauchy kernel function.
  return x=ph, y=ps, z=$\widehat{f}(\phi,\ \psi)$.
end function
begin
1 given the number of grid, n_1.
2 choose random values for bandwidth, $0 \leq h_1,\ h_2 \leq 1$.
3 change the value of the angle $(\phi_i,\ \psi_i)\ i = 1,\ 2,\ .....,\ m$ to the radians mesure.
4 lims=$\{-\pi,\ \pi,\ -\pi,\ \pi\}$.
5 fhat=f_hat(phi, psi, h1, h2, n_1, lims).
6 go to the function f_hat.
end

3. Application

The approach that we have suggested in Algorithm 1 is applied to two types of data (simulated data and real data) in order to evaluate the performance this approach and to examine the effect of two factors. The first factor we would to examine is the smoothing parameters while the second factor is the kernel function. In Section 3.1 we will present the affect of these factors under circular simulated data, while in Section 3.2 we will show the influence of these factors under real data.

3.1 Simulated Data

Let $(\boldsymbol{\phi_1},\ \boldsymbol{\psi_1}), (\boldsymbol{\phi_2}, \boldsymbol{\psi_2}), \ldots, (\boldsymbol{\phi_m}, \boldsymbol{\psi_m})$ be bivariate circular data which have been generated from von Mises distribution vM$(\mathbf{0}, \mathbf{10})$, and let $\boldsymbol{h_1}$ and $\boldsymbol{h_2}$ be the smoothing parameters. The approach that we have suggested is applied to bivariate circular data, which have been generated from von Mises distribution vM$(\boldsymbol{\mu}, \boldsymbol{\kappa})$ with different kernel functions (Wrapped Cauchy distribution and Wrapped Normal distribution) and the results are presented in Figure 1.

As we can see in Figure 1, with different kernel functions the values of the smoothing parameters $\boldsymbol{h_1}$ and $\boldsymbol{h_2}$ have a strong influence on the shape of the kernel density estimation. While, the kernel functions have also an effect on the pattern of the kernel density estimation.

International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol.4 Issue 9, September-2020, Pages: 18-25

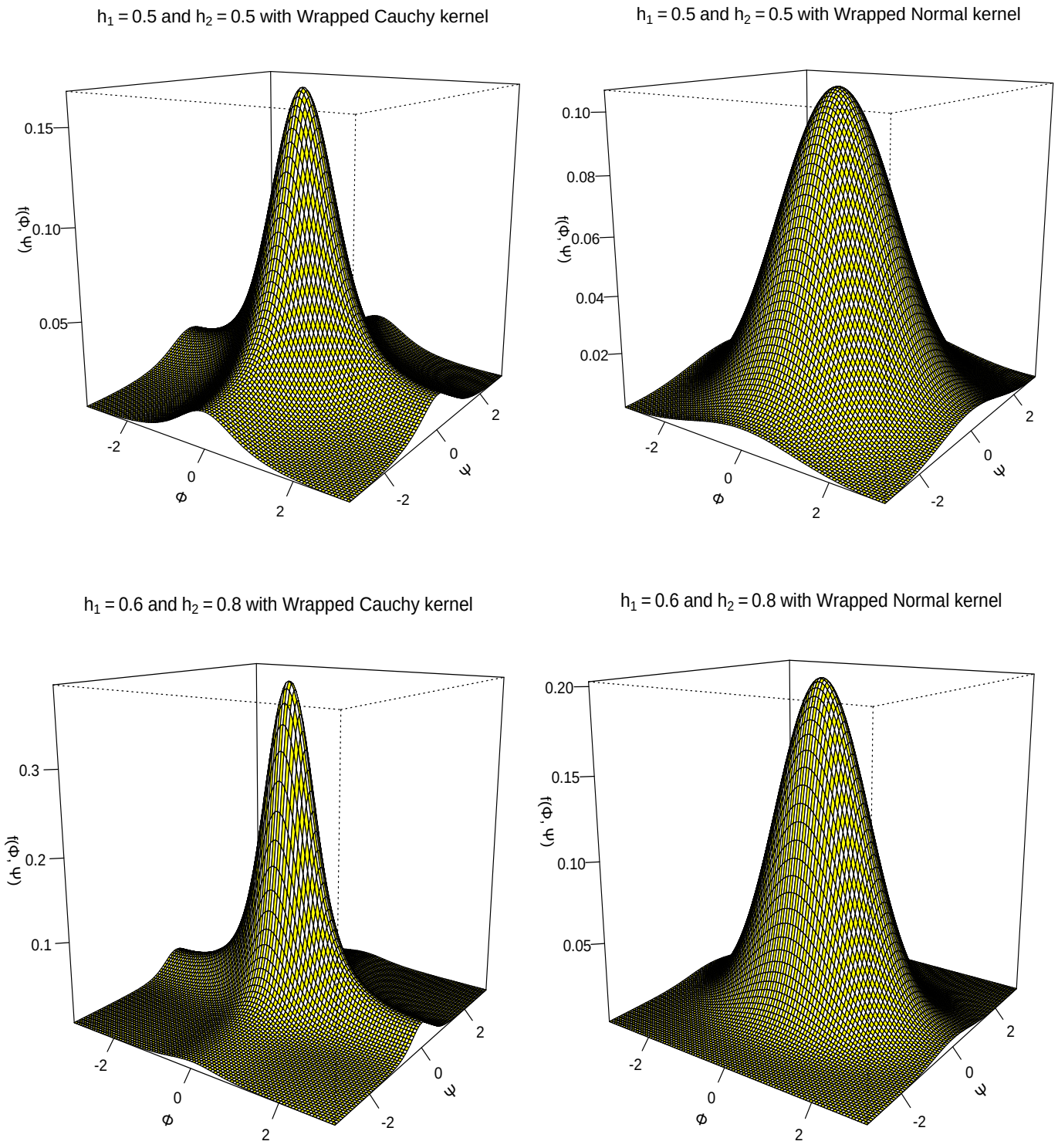

Figure 1: Kernel density estimation for bivariate circular data which have been generated from von Mises distribution vM$(\mathbf{0}, \mathbf{10})$. The estimation has been done with different kernel functions (Wrapped Cauchy distribution and Wrapped Normal distribution), and different smoothing parameters $\boldsymbol{h_1}$and $\boldsymbol{h_2}$.

3.2 Protein Data

A protein is a complex molecule in food that play many critical votes in our bodies. Most of the work in cell can be done by the

proteins which are required for the structure and function of the body's origins and tissues. Each protein can be defined as an ordered sequence of amino acids. These amino acids are connected together by peptide bonds to form a protein backbone. Protein can be represented by a sequence of atoms $N_1 - C_1^{\alpha} - C_1 - N_2 - C_2^{\alpha} - C_2 - \cdots - N_m - C_m^{\alpha} - C_m$. The confirmation of the protein backbone can be described by three angles $(\phi_i, \psi_i, \omega_i)$ which are called the dihedral angles (torsion angles). Please note that these angles are not between the intersection of two straight lines but they are the internal angles of protein backbone at which to adjacent planes meet, where a plane is a two-dimensional surface (see Figure 2). The values of angles ϕ_i and ψ_i are defined to be in the interval $[-\pi, \pi)$, while the value of angle ω_i is generally fixed about zero, for this reason only the angles (ϕ_i, ψ_i), which can be represented as a point on the torus, are helpful in understanding the protein backbone. For a specific protein, if the dihedral angles (ϕ_i, ψ_i) are known then the structure of the protein is almost determined, where each pair of angles can be viewed as a point on the torus.

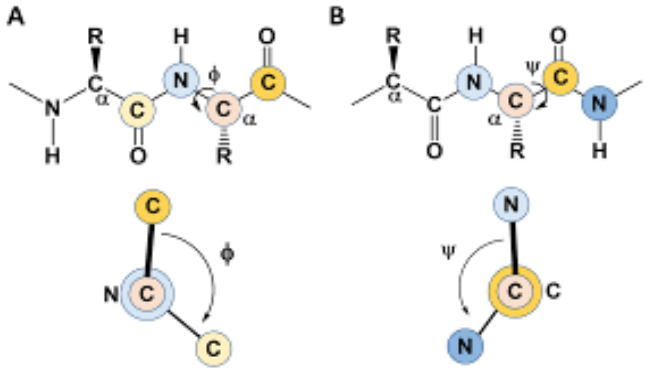

Figure 2: This figure shows dihedral angles (ϕ, ψ) in the protein backbone.

The approach that we have suggested is applied to real data, which have been downloaded from the Protein Data Bank (PDB) https://www.rcsb.org/. From PDB we select two types of proteins 1QQ5 and 5ZXD that published (uploaded to the repository) from 2015 to May 2019 (inclusive), and have less than 1.5^{o}A in resolution. These proteins together consist of residuals and for each case, we are given the following:

1- Dihedral angles ϕ_i: this angles describe the rotation of the protein chain around the bond $N_i - C_i^{\alpha}$, where $\phi_i \in [-\pi, \pi]$.
2- Dihedral angles ψ_i: this angles describe the rotation of the protein chain around the bond $C_i^{\alpha} - C_i$, where $\psi_i \in [-\pi, \pi]$.
3- Amino acids: the type of amino acid which is associated with the dihedral angles in the protein position.
4- Position: the position of amino acid in the protein sequence, where in each position there is a specific dihedral angles (ϕ_i, ψ_i) for each amino acid.

Let $(\boldsymbol{\phi_1}, \boldsymbol{\psi_1}), (\boldsymbol{\phi_2}, \boldsymbol{\psi_2}), \ldots, (\boldsymbol{\phi_m}, \boldsymbol{\psi_m})$ be dihedral angles which belong to proteins 1QQ5 and 5ZXD, and let $\boldsymbol{h_1} = \mathbf{0.5}$ and $\boldsymbol{h_2} = \mathbf{0.5}$ be the smoothing parameters which have been choosing randomly. The approach that we have suggested is applied to the dihedral angles for this type of protein and the results are presented in Figures 3 and 4. The top panels of these figures show the probability density function for the proteins 1QQ5 and 5ZXD under Wrapped Cauchy distribution and Wrapped Normal distribution as a kernel function. While the bottom panels show the contour plot for the joint probability density function for the dihedral angles $(\boldsymbol{\phi}, \boldsymbol{\psi})$.

In the practical, we notice that with different kernel functions (Wrapped Cauchy distribution and Wrapped Normal distribution) the values of the smoothing parameters $\boldsymbol{h_1}$ and $\boldsymbol{h_2}$ have an influence on the shape of the kernel density estimation. Moreover, as we can see in Figures 3 and 4 the kernel functions have also an effect on the pattern of the kernel density estimation.

International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol.4 Issue 9, September-2020, Pages: 18-25

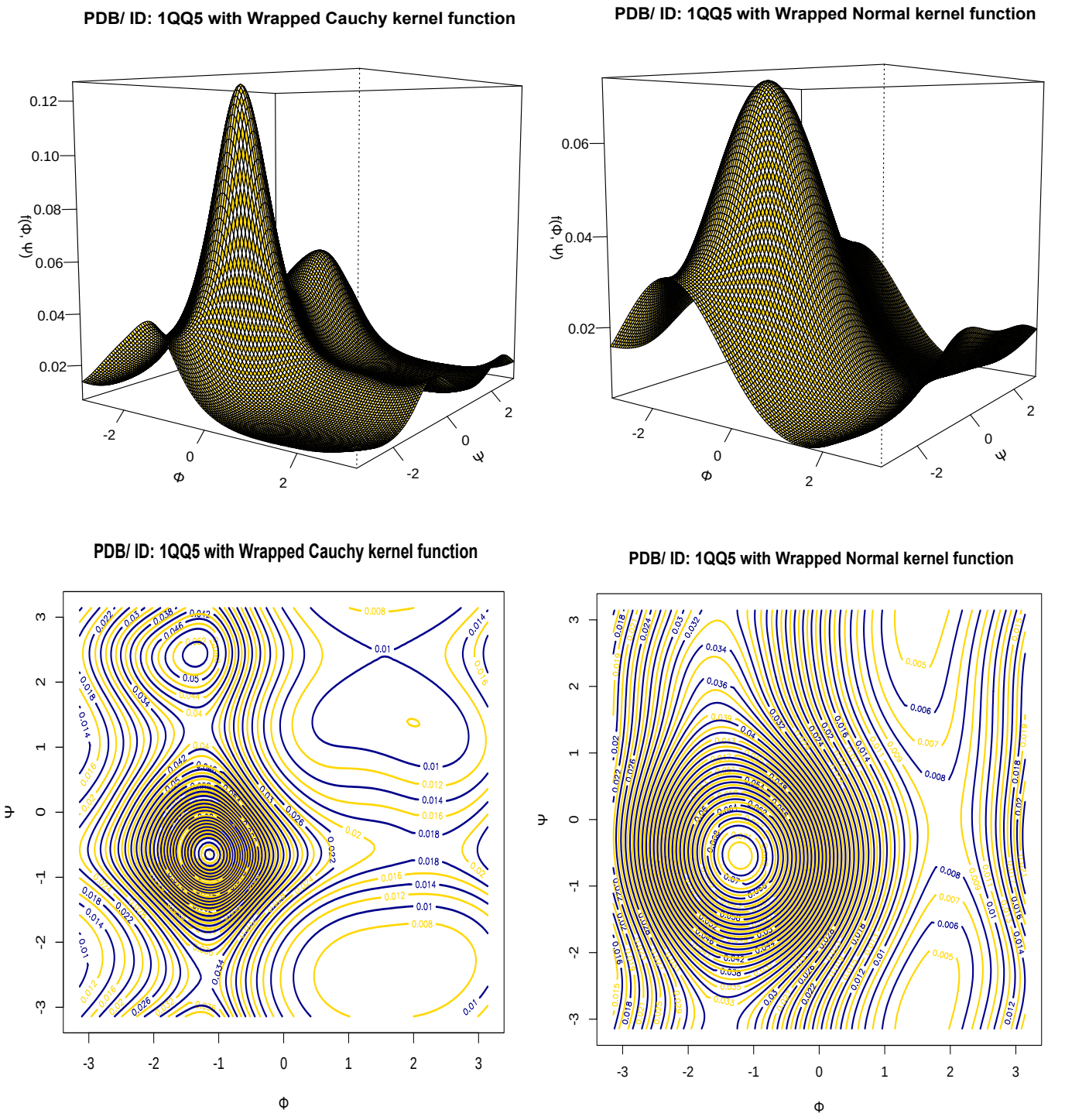

Figure 3: Top panels show the kernel density estimation for dihedral angles $(\boldsymbol{\phi}, \boldsymbol{\psi})$ which belong to protein 1QQ5. Bottom panels show contour plot for the dihedral angles $(\boldsymbol{\phi}, \boldsymbol{\psi})$ which belong to the same protein.

International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol.4 Issue 9, September-2020, Pages: 18-25

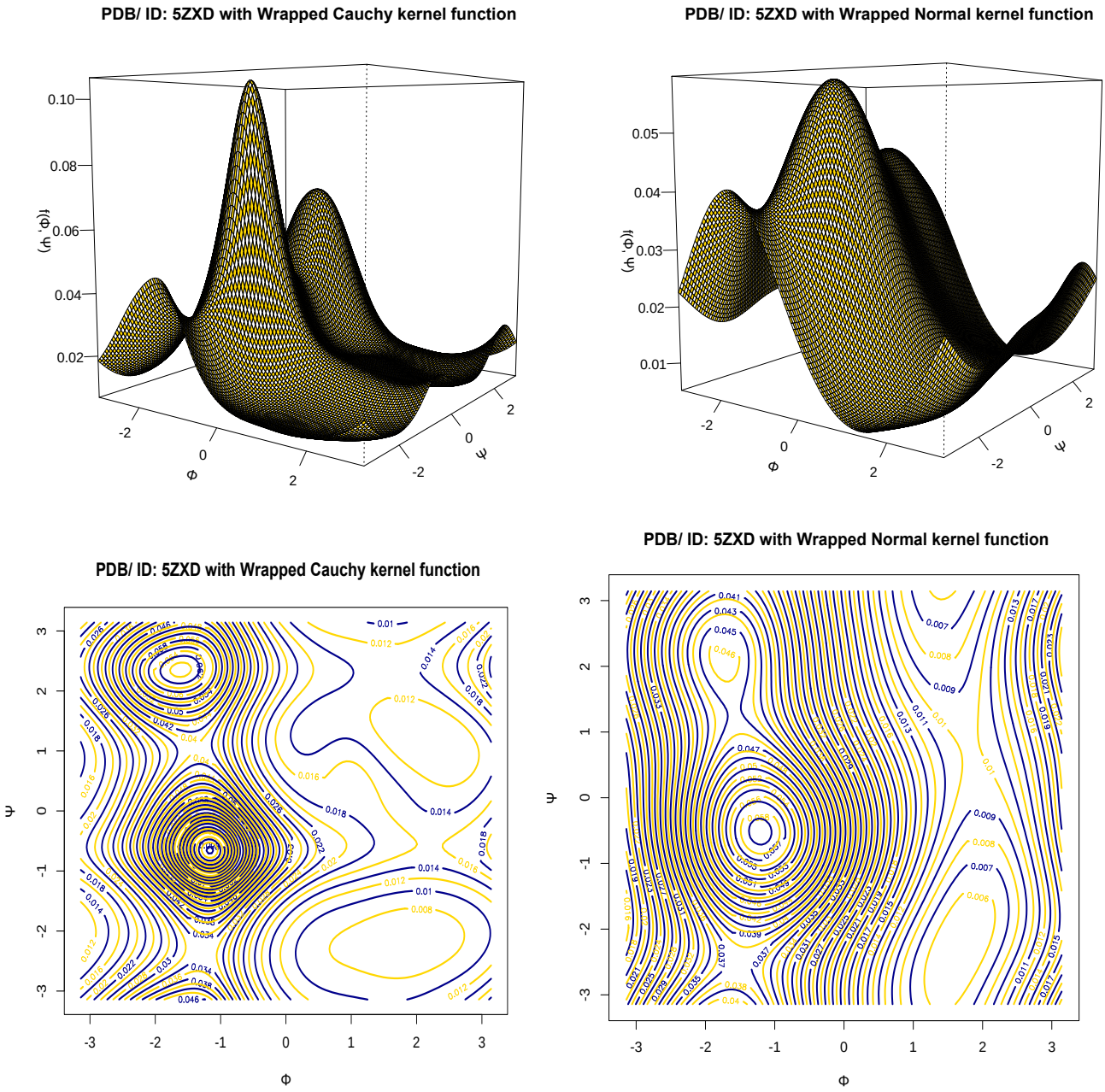

Figure 4: Top panels show the kernel density estimation for dihedral angles $(\boldsymbol{\phi}, \boldsymbol{\psi})$ which belong to protein 5ZXD. Bottom panels show contour plot for the dihedral angles $(\boldsymbol{\phi}, \boldsymbol{\psi})$ which belong to the same protein.

4. Conclusion

In this paper, we developed an approach to estimate the probability density function on the torus for bivariate circular data, this approach may help researchers to explain the pattern of some molecules such as proteins and amino acids. Kernel density estimation with different kernel functions (Wrapped Cauchy distribution and Wrapped Normal distribution) are used to construct this procedure. Moreover, we evaluate the performance of the procedure that we have suggested with simulated data and protein data in order to examine the effect of two factors: smoothing parameters and kernel functions. In the practical, we noticed that with different kernel functions (Wrapped Cauchy distribution and Wrapped Normal distribution) the values of the smoothing parameters and the type of the kernel functions have an influence on the pattern of the kernel density estimation.

References

[1] Abushilah, S. F.[2019], *'Clustering methodology for bivariate circular data with application to protein dihedral angles'*, PHD thesis, University of Leeds.

[2] Chakraborty, S. and Wong, S.W. [2017], 'BAMBI: An R package for fitting bivariate angular mixture models', *arXiv preprint arXiv: 1708.07804*.

[3] Di Marzio, M., Panzera, A. and Taylor, C. C. [2011], 'Kernel density estimation on the torus', *Journal of Statistical Planning and Inference* 141(6), 2156–2173.

[4] Fisher, N. [1989], 'Smoothing a sample of circular data', *Journal of Structural Geology* 11(6), 775–778.

[5] Hall, P., Watson, G. and Cabrera, J. [1987], 'Kernel density estimation with spherical data', *Biometrika* 74(4), 751–762.

[6] Hornik, K. and Grün, B. [2014], 'movMF: An R package for fitting mixtures of von mises-fisher distributions', *Journal of Statistical Software* 58(10), 1–31.

[7] Mardia, K. V. and Jupp, P. E. [2000], *Directional Statistics*, John Wiley & Sons.

[8] Silverman, B. [1986], *Density Estimation for Statistics and Data Analysis*, Chapman & Hall/CRC Press.

[9] Sasaki, H., Noh, Y.-K. and Sugiyama, M. [2015], Direct density-derivative estimation and its application in kl-divergence approximation, *in* 'Artificial Intelligence and Statistics', pp. 809–818.

[10] Taylor, C. C., Mardia, K. V., Di Marzio, M. and Panzera, A. [2012], 'Validating protein structure using kernel density estimates', *Journal of Applied Statistics* 39(11), 2379–2388.

[11] Wand, M. P. and Jones, M. C. [1995], Kernel Smoothing, Chapman & Hall/CRC Press.