Journal of Multivariate Analysis 143 (2016) 374–382

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

A multivariate circular distribution with applications to the protein structure prediction problem

Sungsu Kim [a,*], Ashis SenGupta [b], Barry C. Arnold [c]

[a] Worcester Polytechnic Institute, United States
[b] Indian Statistical Institute, India
[c] University of California, Riverside, United States

ARTICLE INFO

Article history:
Available online 22 October 2015

AMS subject classification:
62J99

Keywords:
Asymmetric generalized von Mises distribution
Bioinformatics
Marginally specified distribution
Multivariate circular distribution
Test of independence

ABSTRACT

The protein structure prediction problem is considered to be the holy grail of bioinformatics, and circular variables in protein structure problem are ubiquitous. For example, conformational angles appear in γ turns, α helices, and β sheets. It is well known that dihedral angles (ϕ and ψ) together with ω (torsion angle of the peptide bond) and χ (torsion angle of the side chain) are considered to be important for protein structure prediction since they define the entire conformation of a protein. In order to study k conformational angles, we need a k-variate angular distribution. In this paper, we propose a multivariate circular distribution and inferential methods, which could be useful for jointly modeling those circular variables of interest. Our proposed family of k-variate circular distributions and testing methods are applied to trivariate circular data set arising from γ turns consisting of Glycine–Phenylalanine–Threonine sequences. We have shown that there is a three-way dependent relationship between the ϕ, ψ and χ, and that the side chain angles are relevant to the relationship between dihedral angles for the given sequence. The proposed model was compared with two existing multivariate circular models using bivariate and trivariate circular data sets.

1. Introduction

Proteins are linear polymers of the 20 common amino acid residues. Different combinations and different lengths constitute different proteins. An amino acid is made up of a central carbon atom called the C-alpha to which is bonded an amine group, a carboxyl group, a hydrogen atom and a side chain. The torsion angle of the bond between N of the amine group and C-alpha is denoted by ϕ and that between C of carboxyl group and C-alpha is denoted by ψ, while that of the side chain is designated by χ. Condensation of the carboxyl group of one amino acid and the amine group of another results in formation of the peptide bond, whose torsion angle is called ω. The protein structure prediction problem is considered to be the holy grail in bioinformatics. One way to tackle this problem is homology modeling based on the idea that similar amino acid residue sequences may have similar protein structures. The problem states "How to predict the exact three dimensional structure of a protein from its one dimensional amino acid residue sequence?" The covalent structure of a polypeptide chain is not sufficient to determine its three dimensional structure, due to the possibility of different rotations about the many covalent bonds. Three-dimensional structures that differ only in this way are referred to as conformations.

* Corresponding author.
E-mail address: dr.sungsu@gmail.com (S. Kim).

http://dx.doi.org/10.1016/j.jmva.2015.09.024
0047-259X/

The conformation of the polypeptide backbone is defined by the torsion angles ϕ, ψ, and ω of each residue. The study of these angles in proteins is relevant in the prediction of the three dimensional structure from the sequence of its amino acid residues. One of the common secondary structures of the protein is known as the γ turn, which implies a trivariate circular distribution of the dihedral angles ϕ, ψ and χ. It is important to investigate if the side chain angles χ are relevant to the relationship between the main chain angles, ϕ and ψ [2]. In this paper, we illustrate our methods using 334 γ turns consisting of Glycine–Phenylalanine–Threonine sequences.

When studying multi-dimensional circular random variables, it is often difficult to directly visualize the surfaces pertaining to their multivariate densities. It is often the case that it is easier to introspect about the nature of the univariate circular marginal and conditional circular distributions, and use them to build a suitable multivariate circular model. One of our motivations in this paper is to describe the probability distribution on the hypertorus. We propose a family of k-variate circular models with specified marginals, which is equivalent to adopting Sklar's theorem [15] of the theory of copulas [10]. The new family of distributions can be considered as a multivariate extension of the bivariate family of circular distributions given in Wehrly and Johnson [16]. One attractive property of the new model is that one can study k circular variables, where a given set of $k - m$ circular variables is mutually independent. A multivariate circular distribution as an extension to the von Mises (vM) distribution was presented in Mardia et al. [9]. They proposed estimation and hypothesis testing methods and illustrated the utility of their model using protein data of dihedral angles in γ turns. Advantages of our model over their model are discussed in Section 4. Another multivariate circular distribution was proposed by Fernandez-Duran and Gregorio-Dominguez [4], which is a multivariate extension of the univariate model proposed in Fernandez-Duran [3]. A method of constructing bivariate joint circular density from prescribed conditionals can be found in the Appendix.

In the next section, we propose a new multivariate circular distribution and present some of its properties, and statistical inferential methods. In Section 3, we compare our proposed model with the multiple nonnegative trigonometric sums (MNTS) model [4] and Mardia's model [9] using two real data sets; one bivariate and one trivariate cases. In Section 4, we discuss advantages of the new model over the Mardia's model, along with concluding remarks.

2. Methods

2.1. k-variate circular distribution

Wehrly and Johnson [16] generated a family of bivariate circular distributions with the joint pdf of the form

$$f(\theta_1, \theta_2) = 2\pi \cdot g\left[2\pi\{F_1(\theta_1) \pm F_2(\theta_2)\}\right] f_1(\theta_1) f_2(\theta_2), \tag{2.1}$$

where $-\pi < \theta_1, \theta_2 \leq \pi$, g, f_1 and f_2 are densities on the circle, and F_1 and F_2 are the distribution functions of f_1 and f_2, respectively. The density (2.1) has f_1 and f_2 as its specified marginal densities. The model can be used with any desired marginal densities. For example, Shieh and Johnson [13] studied distributions and inferential questions regarding the family of bivariate distributions of the form (2.1), with von Mises (vM) marginals. Shieh et al. [14] studied the family of bivariate distributions of the form (2.1) with generalized von Mises (GvM) marginals, together with associated inferences. In both papers, a vM distribution for $g(\cdot)$ was assumed.

We propose a family of multivariate extensions of the Wehrly and Johnson [16] model in the following theorem.

Theorem 2.1. (a) *The following is a k-variate circular density function*

$$f(\underline{\theta}) = (2\pi)^m \prod_{j=1}^{m} \left\{ g_j\left(2\pi \sum_{i=j}^{k} F_i(\theta_i)\right)\right\} \prod_{i=1}^{k} f_i(\theta_i), \tag{2.2}$$

where $\underline{\theta} \in (-\pi, \pi]^k$, $1 \leq m \leq k-1$, *the* f_i' *s and* g_j' *s are circular densities and* $F_i(\theta_i) = \int_{-\pi}^{\theta_i} f_i(\eta_i) d\eta_i$ *for* $i = 1, \ldots, k$ *and* $j = 1, \ldots, m$. *By convention,* $2\pi \sum_{i=j}^{k} F_i(\theta_i)$ *in the argument of each* $g_j(\cdot)$ *is interpreted as modulo* 2π.

(b) *The density has the property that,* $\theta_{m+1}, \ldots, \theta_k$ *are jointly independent.*

(c) *For* $j = 1, \ldots, m$, *the conditional density of* θ_j *given* $\theta_{j+1}, \ldots, \theta_k$ *is given by*

$$f(\theta_j|\theta_{j+1}, \ldots, \theta_k) = 2\pi \cdot g_j\left(2\pi \sum_{i=j}^{k} F_i(\theta_i)\right) f_j(\theta_j). \tag{2.3}$$

Assuming that g_j *is a von Mises density,* (2.3) *has one parameter, the concentration parameter* κ, *for testing conditional independence of* θ_j *given* $\theta_{j+1}, \ldots, \theta_k$.

(d) *For* $j = 1, \ldots, m$, *the conditional density of* θ_j *given* $\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k$ *is given by*

$$f(\theta_j|\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k) = (2\pi)^j \prod_{s=1}^{j} g_s\left(2\pi \sum_{i=s}^{k} F_i(\theta_i)\right) f_j(\theta_j). \tag{2.4}$$