ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

Dairesel Veri Analizinde İki-Tepeli ve Karma İstatistiksel Dağılımlar

Bi-Modal and Mixture Distributions in Circular Data Analysis

Muhammet Burak KILIÇ[a]

[a]İstatistik Bölümü,
Mehmet Akif Ersoy Üniversitesi
Fen-Edebiyat Fakültesi,
Burdur

Geliş Tarihi/*Received:* 14.03.2016
Kabul Tarihi/*Accepted:* 21.06.2016

Yazışma Adresi/*Correspondence:*
Muhammet Burak KILIÇ
Mehmet Akif Ersoy Üniversitesi
Fen-Edebiyat Fakültesi,
İstatistik Bölümü, Burdur,
TÜRKİYE/TURKEY
mburak@mehmetakif.edu.tr

ÖZET Amaç: Dairesel istatistik, açı olarak gözlenen verileri birim daire üzerinde analiz eden özel bir alandır. Çevre, biyoloji veya tıp gibi çok çeşitli çalışmalarda dairesel (açısal) veriler araştırmanın önemli bir bölümünü oluşturmaktadır. Örneğin, proteinlerdeki dihedral açılardan faydalanılarak proteinlerin ikincil yapısının belirlenmesi ya da geometrik morfolojide eklemi oluşturan kemikler arasındaki açılarla yürüyüş bozuklukları gibi fiziksel bozuklukların belirlenmesi, ya da su kaplumbağalarının yumurtalarını bıraktıktan sonra sahile dönüş yönleri ile kumsalların planlanması bu alanın önemli uygulamalarındandır. Bu çalışmalarda doğrusal istatistik yöntemlerin kullanılması dairesel verilerin geometrik şeklinden dolayı yanıltıcı sonuçlara yol açmaktadır. Bundan dolayı bu tür açısal verileri analiz ederken, dairesel istatistik yöntemleri kullanılmalıdır. Bu çalışmanın amacı, dairesel veri analizinde iki-tepeli ve karma dağılımları karşılaştırmaktır. **Gereç ve Yöntemler:** Bu çalışmada, iki-tepeli karma von Mises, sarmal Normal ve sarmal Cauchy dağılımları ile genelleştirilmiş von Mises dağılımları kullanılmış ve tahmin edicileri elde etmek için iteratif yöntemler kullanılmıştır. İteratif yöntemler R programında uygulanmış ve dairesel dağılımların parametre tahmini için R kodları verilmiştir. Bununla birlikte, bu dairesel dağılımlar, proteinlerdeki dihedral açıları ve su kaplumbağalarının kumsaldan ayrılış yönleri verileriyle incelenmiş ve model seçimi Akaike ve Bayesci bilgi kriterleri kullanılarak yapılmıştır. **Bulgular:** Proteinlerdeki dihedral açılar için, en iyi uyumu iki-tepeli sarmal Cauchy dağılımı vermiştir. Su kaplumbağalarının dönme açıları için, en iyi uyumu genelleştirilmiş von Mises ve iki-tepeli karma von Mises dağılımları vermiştir. **Sonuç:** Dairesel verilerde bir veya birden fazla tepe noktasında aşırı bir yoğunlaşma gözlendiğinde, iki-tepeli karma von Mises ve genelleştirilmiş von Mises dağılımları, modelleme olarak tercih edilmeyebilir. Dairesel verilerin tepe noktalarında aşırı bir yoğunlaşma gözlenmediğinde ise genelleştirilmiş von Mises dağılımları dairesel verileri modellemek için önerilebilir.

Anahtar Kelimeler: Dairesel veri analizi; dairesel dağılımlar; açısal veri

ABSTRACT Objective: Circular statistics is a special area which is analyzed by observed angular data on the unit circle. In many various studies, such as environment, biology or medicine, circular (angular) data is an important part of the research. For illustration, to determine the secondary structure of the proteins by utilizing dihedral angles or to asses physical disorders such as gait disturbances between the bones in the geometric morphology or the organization of the beach after leaving the eggs of sea turtles, are the important applications of this area. The uses of linear statistical methods in this area lead to misleading results because of the geometric shape of the circular data. Therefore, when it is analyzing such angular data, circular statistical methods should be used. The objective of this study is compared with the bi-modal and mixture distributions in circular data analysis. **Material and Methods:** The bi-modal mixture von Mises, wrapped Normal, wrapped Cauchy and the generalisations of von Mises distributions were used and it was performed by iterative methods to obtain maximum likelihood estimators. These iterative methods were applied in R programming and the R codes were given for the circular distribution of the parameter estimation. These distributions were examined for analyzing dihedral angles in proteins and turtles rotations, and model selection was performed by using Akaike and Bayesian information criteria. **Results:** For dihedral angles in protein, two mixture wrapped Cauchy distribution was given the better fit. For turtle rotations, the generalizations of von Mises distribution and two mixture von Mises distribution were given the better fit. **Conclusion:** If there is observed an excessive concentration in one or more modes in analyzing circular data, the bi-modal mixture von Mises and the generalisations of von Mises distribution for modeling may not be preferred. If there is not observed an excessive concentration in analyzing the circular data, then the generalised von Mises distributions may be proposed for modeling the circular data.

Key Words: Circular data analysis; circular distributions; angular data

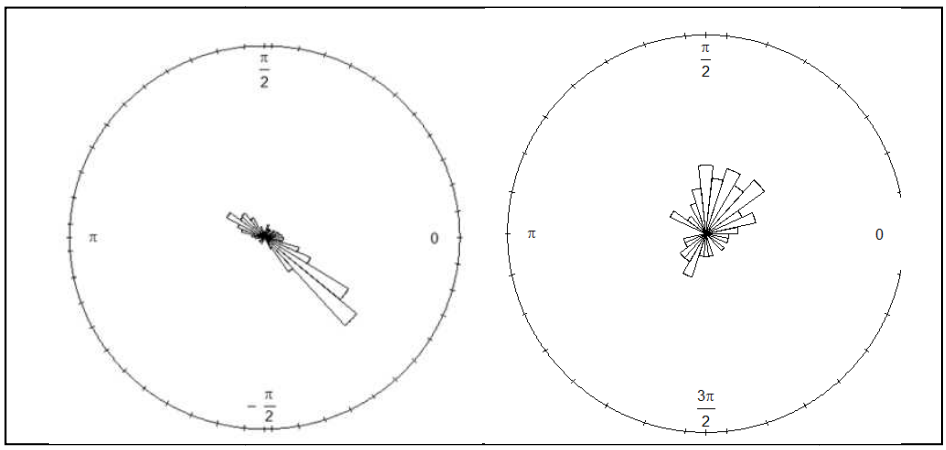
doi: 10.5336/biostatic.2016-51329

Turkiye Klinikleri J Biostat 2016;8(2):133-42

Son yıllarda, özellikle biyoloji, tıp, çevre, gometrik morfoloji veya astronomi alanlarında yapılan araştırmalarda dairesel (açısal) veriler önemli bir yer tutar. Örneğin bioinformatik alanında protein moleküllerinin yapısının belirlenmesinde, açısal değişkenlerin modellenmesi önemlidir. Bu açısal değişkenlerde çoğunlukla iki-tepeli olarak gözlenmektedir.[1] Örnek olarak Dihedral Angle Secondary Structure Database[1] (DASSD) protein veri tabanındaki Alanin amino asitler dizisi (Ala-Ala-Ala) için dihedral açısı ψ'nin iki tepe noktasına sahip olduğu gözlenmektedir (Şekil 1a). Dairesel verilerin bir diğer önemli uygulaması ise biyoloji ve çevrebilimi ile ilgili araştırmalarda ortaya çıkmaktadır. Örneğin biyolojik araştırmalarda, yumurtadan yeni çıkan su kaplumbağalarının hangi yöne gideceklerine dair bilgiler dairesel (açısal) veri analizi ile elde edilmektedir. Bu tür bilgiler, su kaplumbağaların yumurtalarını bıraktıkları kumsalların daha çevre dostu bir şekilde planlanmasını sağlar. Diğer bir örnek, Gould tarafından gözlenmiş olan 76 su kaplumbağasının yumurtalarını kumsala bıraktıktan sonra kumsaldan dönüş açıları θ iki tepeli olarak gözlenmiştir (Şekil 1b).[2] Burada su kaplumbağalarının büyük bir çoğunluğunun belli bir yöne gittiği görülürken, çok az bir kısmının farklı bir yöne gittiği görülmektedir (Şekil 1b). Bu örneklerden açıkça anlaşıldığı gibi, iki tepeli olarak gözlenen dairesel veriler için tek tepeli dairesel dağılımlarla modelleme yapmak yerine karma veya iki tepeli dairesel modelleme yapılarına ihtiyaç duyulmaktadır.

İki tepeli dairesel verileri modellemek için çoğunlukla iki yaklaşım tercih edilir. İlk olarak tek tepeli dairesel dağılımların karmaları alınarak, iki-tepeli karma dairesel dağılımların kullanılmasıdır. Özellikle, literatürde iki-tepeli karma von Mises dağılımlarının kullanıldığı bazı önemli çalışmalar mevcuttur.[3-5] İki tepeli karma dairesel dağılımlara alternatif bir yaklaşım ise genelleştirilmiş von Mises dairesel dağılımlarıdır. Genelleştirilmiş von Mises dağılımlarının en önemli özellikleri, tek tepeli, iki tepeli simetrik veya asimetrik dairesel verileri modellemek için kullanılmasıdır.[6-9] Ancak literatürde, iki tepeli dairesel veriler için, hangi durumlarda, hangi dairesel dağılımların kullanılacağı hakkında bir belirsizlik vardır.

Bu çalışmada iki tepeli karma von Mises, sarmal normal ve sarmal Cauchy modelleri ile genelleştirilmiş von Mises ve asimetrik genelleştirilmiş von Mises modellerinin gerçek dairesel veriler üzerinde model seçim kriterlerleriyle karşılaştırılması amaçlanmaktadır. Makalenin ikinci



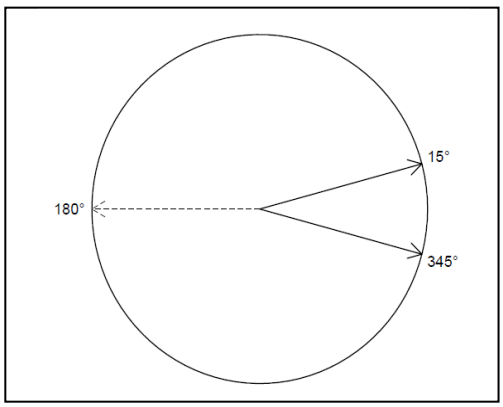
ŞEKİL 1: (a) ψ açılarının gül diyagramı. **(b)** θ açılarının gül diyagramı.

bölümünde, ilk olarak dairesel verilerin tanımlayıcı istatistikleri kısaca özetlenmiştir. Daha sonra tek tepeli, iki tepeli karma ve genelleştirilmiş von Mises dairesel dağılımlarına yer verilmiş ve dairesel modellerin parametre tahmin edicilerini elde etmek için en çok olabilirlik yöntemine bağlı olarak, iteratif yöntemlere ait komutlar R programında yazılmış, ilgili kodlar verilmiştir. Son bölümünde ise proteinlerdeki dihedral açısı ve su kaplumbağalarının kumsaldan dönüş açıları verileriyle dairesel modelleme yapılmış olup modeller Akaike ve Bayesci bilgi kriterleriyle karşılaştırılmıştır. Elde edilen bulgular sonuç kısmında tartışılmıştır.

GEREÇ VE YÖNTEMLER

DAİRESEL İSTATİSTİK

Dairesel veriler, bir daire üzerinde derece veya radyan birimleri ile tanımlanabilen açısal gözlemlerdir ve tanım aralıkları sınırsız bir doğrudan ziyade [0, 2π) veya [-π, π)'dir. Gözlem değerleri daire üzerinde derece veya radyan birimleri ile ifade edildiğinden tanımlayıcı istatistikleri trigonometrik hesaplamalara dayanmaktadır. Dolayısıyla dairesel verilerin analizi, doğrusal verilerde kullanılan standart analiz yöntemlerinden farklıdır. Örneğin, Şekil 2'de gösterildiği gibi, 15^0 ve 345^0 açıları iki dairesel veri olsun. Bu değerlerin aritmetik ortalaması 180°'dir. Dairesel ortalaması ise iki vektörün toplamının yönü, yani bileşke vektör yönü 0^0 olur. Dolayısıyla basit örneklem ortalaması gibi doğrusal formüller dairesel veri analizi için kullanılmamalıdır.



ŞEKİL 2: Daire üzerindeki dairesel (açısal) veri gösterimi.

Doğrusal tanımlayıcı istatistiklere paralel olarak dairesel tanımlayıcı istatistiklerde mevcuttur. Dairesel ortalama bulunurken, gözlemlerin vektörel özelliklerinden faydalanılır. $\theta_i \epsilon [0, 2\pi)$, $i = 1,2,\dots,n$ gözlenmiş dairesel veriler olsun. Her bir gözlem için, dikdörtgensel dönüşüm uygulandığında, $(\cos\theta_i, \sin\theta_i)$, $i = 1,2,\dots,n$ elde edilir. Burada, her bir gözlemin vektörel bileşenleri toplandığında, bileşke vektör aşağıdaki gibi tanımlanır.

$$\boldsymbol{R} = \left(\sum_{i=1}^{n} \cos\theta_i \, , \sum_{i=1}^{n} \sin\theta_i \right) = (C, S)$$

Bileşke vektörün uzunluğu $R = \|\boldsymbol{R}\| = \sqrt{C^2 + S^2}$, formülünden bulunur. Bileşke vektörün yönü ise dairesel ortalamayı gösterir. Ayrıca dairesel ortalama $\bar{\theta}$ aşağıdaki ifadelerden de elde edilir.

$$\bar{\theta} = \arctan^*(S/C) = \begin{cases} \arctan(S/C), & \text{if } C > 0, S \geq 0, \\ \pi/2, & \text{if } C = 0, S > 0, \\ \arctan(S/C) + \pi, & \text{if } C < 0, \\ \arctan(S/C) + 2\pi, & \text{if } C \geq 0, S < 0, \\ \text{undefined}, & \text{if } C = 0, S = 0. \end{cases}$$

Örnek dairesel varyans ise $V = 1 - \bar{R}$ formülünden hesaplanır. Doğrusal istatistiğe paralel olarak, küçük dairesel varyans da verilerin dağılımının daha yoğun olduğunu gösterir. Örnek dairesel standart sapma ise, $\{-2\log(1-V)\}^{1/2}$ formülü ile elde edilir.

Dairesel veri analizi, tanım aralığının sınırlı ve dairesel ortalama gibi tanımlayıcı istatistiklerinin de hassas olmasından dolayı doğrusal analize göre daha zordur. Yine de dairesel veri analizinde kullanılan önemli istatistiksel metotlar vardır.[10-12]

VON MİSES DAĞILIMI

Dairesel dağılımlar arasında en çok kullanılan dağı-

lım von Mises dağılımıdır. Bu dağılım dairesel normal dağılım olarak da adlandırılmaktadır.[12] θ rassal değişkeni von Mises dağılımına sahip ise olasılık yoğunluk fonksiyonu aşağıdaki şekilde ifade edilir.

$$h_{vM}(\theta) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta-\mu)}, \quad 0 \le \mu < 2\pi, \ \kappa > 0 \qquad (1)$$

Dağılımın dairesel ortalama ve yoğunlaşma parametreleri μ ve κ'dır. $I_0(\kappa)$ ise birinci tür ve sıfır sırasında dönüştürülmüş Bessel fonksiyonudur. Bu fonksiyonun birinci tür ve p.nci sırası ise aşağıdaki şekilde tanımlanır.

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

Birinci tür ve sıfır sırasında dönüştürülmüş Bessel fonksiyonu için bazı polinomsal yaklaşımlarda mevcuttur.[13] Şekil 3'de von Mises olasılık yoğunluk fonksiyonları μ=0⁰ ortalama yön ve farklı yoğunluk parametreleri κ=1,2,7 ve 10 için birim daire üzerinde gösterilmiştir.

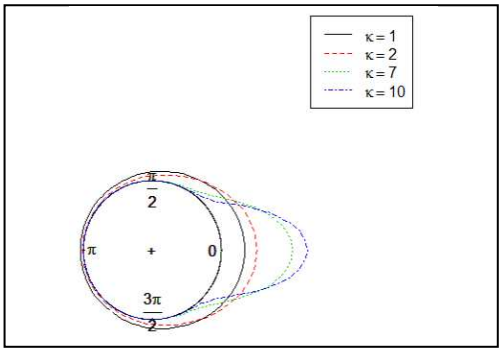
SARMAL DAĞILIM AİLESİ

Dairesel dağılımlar arasında kullanılan diğer dağılımlar ise sarmal dağılım ailesidir. Sarmal yaklaşım, doğrunun daire çevresinde sarılarak elde edilen bir yaklaşımdır. Örneğin, Y rassal değişkeni, $\mathbb{R}$'de tanımlı ve olasılık yoğunluk fonksiyonu g(y) olsun. Bu durumda daire üzerindeki rassal değişken aşağıdaki şekilde ifade edilir.

$$\theta = Y \bmod (2\pi)$$

Sarmal dağılımların olasılık yoğunluk fonksiyonu ise

$$h(\theta) = \sum_{k=-\infty}^{\infty} g(\theta + 2\pi k)$$



ŞEKİL 3: von Mises olasılık yoğunluk fonksiyonlarının μ=0⁰ ortalama yön ve farklı yoğunluk parametreleri κ=1,2,7 ve 10 için dairesel gösterimi.

ile tanımlanır.[11,12] Sarmal dairesel dağılımlar arasında en çok kullanılan dağılımlar sarmal normal ve sarmal Cauchy dağılımlardır.

Sarmal Normal Dağılım

Sarmal normal dağılım (SN) normal dağılımın $N(\mu, \sigma^2)$ bir daire etrafında sarılmasıyla elde edilir ve von Mises dağılımı vM (μ,κ) gibi dairesel normal dağılım olarak adlandırılır. Bu dağılımın olasılık yoğunluk fonksiyonu ise aşağıdaki şekilde tanımlanır.

$$h_{sN}(\theta) = \frac{1}{\sigma\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} \exp\left[\frac{-(\theta-\mu+2\pi k)^2}{2\sigma^2}\right], 0 \le \theta < 2\pi \qquad (2)$$

Bu olasılık yoğunluk fonksiyonun bir diğer formu ise

$$h_{sN}(\theta) = \frac{1}{2\pi}\left(1 + 2\sum_{p=1}^{\infty} \rho^{p^2} \cos p(\theta-\mu)\right)$$

ve $\rho = e^{-\sigma^2/2}$ olarak tanımlanır. Bu form Normal dağılımın karakteristik fonksiyonundan elde edilir.[12] Bu dağılımda tek tepeli ve simetrik bir dağılımdır.

Sarmal Cauchy Dağılım

Sarmal dağılımlar ailesinin diğer bir önemli dağılımı ise sarmal Cauchy dağılımıdır. Benzer şekilde Cauchy dağılımının daire etrafında sarılarak elde edilen bir dağılımdır. Bu dağılımın olasılık yoğunluk fonksiyonu aşağıdaki şekilde ifade edilir.

$$h_{sC}(\theta) = \frac{1}{2\pi}\left(1 + 2\sum_{p=1}^{\infty} \rho^{p} \cos p(\theta-\mu)\right)$$

$$= \frac{1}{2\pi}\frac{1-\rho^2}{1+\rho^2-2\rho\cos(\theta-\mu)}, \quad 0 \le \mu < 2\pi, \quad 0 \le \rho < 1 \qquad (3)$$

Sarmal Cauchy dağılımının konum parametresi μ, ρ ise yoğunluk parametresidir.[12] Sarmal Cauchy dağılımı tek tepeli ve simetrik bir dağılımdır. Bu dağılım daire etrafında tepe noktası yoğun olan dairesel veriler için önerilir.

İKİ TEPELİ DAİRESEL DAĞILIMLAR

İki tepeli dairesel dağılımlar genel olarak iki şekilde tanımlanabilir. En basit anlamda tek tepeli simetrik dağılımların karmaları alınarak elde edilir. Bu makalede kullanılan iki-tepeli karma olasılık yoğunluk fonksiyonları aşağıdaki gibi gösterilir.

- İki tepeli karma von Mises dağılımı

$$h(\theta_i; p, \mu_1, \kappa_1, \mu_2, \kappa_2) = p h_{vM}(\theta_i; \mu_1, \kappa_1) + (1-p) h_{vM}(\theta_i; \mu_2, \kappa_2)$$

- İki tepeli karma sarmal Normal dağılımı

$$h(\theta_i; p, \mu_1, \rho_1, \mu_2, \rho_2) = p h_{sN}(\theta_i; \mu_1, \rho_1) + (1-p) h_{sN}(\theta_i; \mu_2, \rho_2)$$

- İki tepeli karma sarmal Cauchy dağılımı

$$h(\theta_i; p, \mu_1, \rho_1, \mu_2, \rho_2) = p h_{sC}(\theta_i; \mu_1, \rho_1) + (1-p) h_{sC}(\theta_i; \mu_2, \rho_2)$$

Diğeri ise üstel dağılım ailesine ait olan genelleştirilmiş von Mises dairesel dağılımlarıdır. Bu dağılımların en önemli özelliği tek tepeli, simetrik, asimetrik ve aynı zamanda parametrelere bağlı olarak iki tepeli bir dağılım olmalarıdır.

Genelleştirilmiş von Mises Dağılımı

θ rassal değişkeni genelleştirilmiş von Mises dağılımına sahip olsun. Bu durumda olasılık yoğunluk fonksiyonu

$$h_{gvM}(\theta) = \frac{1}{2\pi c(\delta, \kappa_1, \kappa_2)} \exp(\kappa_1 \cos(\theta - \mu_1) + \kappa_2 \cos 2(\theta - \mu_2)) \qquad (4)$$

(4) ile gösterilir.[6] $\mu_1 \epsilon [0, 2\pi)$, $\mu_2 \epsilon [0, \pi)$, $\delta = \mu_1 - \mu_2$ ve $\kappa_1, \kappa_2 > 0$ şekil parametreleri olarak tanımlanır. Bu dağılım tek tepeli, simetrik ve aynı zamanda iki tepeli bir dağılımdır. Bu dağılımın normalleştirme sabiti ise

$$c(\delta, \kappa_1, \kappa_2) = \frac{1}{2\pi} \int_0^{2\pi} \exp(\kappa_1 \cos(\theta) + \kappa_2 \cos 2(\theta + \delta)\, d\theta \qquad (5)$$

(5) ile ifade edilir.

Üç Parametreli Asimetrik Genelleştirilmiş von Mises Dağılımı

θ rassal değişkeni asimetrik genelleştirilmiş von Mises dağılımına sahipse, olasılık yoğunluk fonksiyonu

$$h_{agvM}(\theta) = \frac{1}{2\pi c\left(\frac{\pi}{4}, \kappa_1, \kappa_2\right)} \exp(\kappa_1 \cos(\theta - \mu) + \kappa_2 \sin 2(\theta - \mu)) \;\; \kappa_1 > 0, \quad \kappa_2 \epsilon [-1, 1] \qquad (6)$$

(6) ile gösterilir[8]. $\mu \epsilon [0, 2\pi)$ konum parametresi, κ_1 yoğunluk parametresi ve κ_2 ise çarpıklık parametresi olarak adlandırılır. Normalleştirme sabiti ise

$$c\left(\frac{\pi}{4}, \kappa_1, \kappa_2\right) = \frac{1}{2\pi} \int_0^{2\pi} \exp(\kappa_1 \cos(\theta) + \kappa_2 \sin 2\left(\theta + \frac{\pi}{4}\right) d\theta \qquad (7)$$

(7) ile ifade edilir. Genelleştirmiş von Mises dağılımlarını analiz ederken normalleştirme sabitlerinin kapalı bir formda olmamasından dolayı, analizi diğer dairesel dağılımlara göre daha zordur.

MODEL PARAMETRELERİNİN TAHMİNİ VE R KODU

Bu çalışmada, dairesel dağılımlarının parametre tahmini en çok olabilirlik yöntemiyle elde edilmiştir. $\theta_i \epsilon [0, 2\pi)$, $i = 1, 2, \ldots, n$ gözlemlerden elde edilmiş dairesel veriler olsun. Dairesel verilerin olasılık yoğunluk fonksiyonu h, $\boldsymbol{\varphi}$ ise parametre kümesi olsun. Bu durumda log olabilirlik fonksiyonu

$$L(\theta_i; \boldsymbol{\varphi}) = \sum_{i=1}^{n} \log\, h(\theta_i;\, \boldsymbol{\varphi}) \qquad (8)$$

(8) ile ifade edilir. Genel olarak, (8) nolu log olabilirlik fonksiyonun birinci türevleri alındıktan sonra parametre tahmin edicileri için kapalı çözümler mevcut olmadığında, tahmin edicileri elde etmek için Newton Raphson, Nelder-Mead veya benzetilmiş tavlama (simulated annealing) iteratif yöntemler kullanılır. Tek tepeli simetrik dağılımların en çok olabilirlik yöntemiyle parametre tahminleri için R programında circular paketi içinde mle.vonmises, mle.wrappednormal, mle.wrappedcauchy fonksiyonları da kullanılabilir. Bununla birlikte maxLik paketi[14] kullanılarak von Mises dağılımının parametre tahminleri için yazılan R kodu:

```
library(circular)
library(maxLik)
theta=rad(dataset) #Dairesel data seti theta, rad: radyan
#mu0 ve k0 başlangıç değerleridir.
#von Mises olasılık yoğunluk fonksiyonu
dvonmises
```

```
#nr - Newton Raphson yöntemini ifade eder.
logLikMix <- function(param) {
mu<- param[1]
kappa <- param[2]
if (kappa < 0)
return(NA)
ll <- log(dvonmises(theta, mu,kappa) )
summary(m1 <- maxLik(logLikMix, start = c(mu = mu0,
kappa = k0),method="nr"))
```

İki tepeli karma dağılımlar için log olabilirlik fonksiyonu

$$L(\theta_i; p, \boldsymbol{\varphi_1}, \boldsymbol{\varphi_2}) = \sum_{i=1}^{n} \mathrm{p} \log\, h(\theta_i; \boldsymbol{\varphi_1}) + (1-\mathrm{p}) \log h(\theta_i;\, \boldsymbol{\varphi_2}) \quad (9)$$

(9) ile tanımlanır. Benzer şekilde, (9) nolu log olabilirlik fonksiyonun birinci türevleri alındıktan sonra parametre tahmin edicileri için kapalı çözümleri mevcut olmadığında, tahmin ediciyi elde etmek için iteratif yöntemlere ihtiyaç duyulmaktadır. İki tepeli karma veya genelleştirilmiş von Mises dağılımının parametre tahminleri için R programında maxLik paketi[15] kullanılarak, başlangıç değerleri p0, mu10, mu20, k10 ve k20 verilerek Newton-Raphson ve Nelder-Mead veya benzetilmiş tavlama (simulated annealing) iteratif yöntemleriyle parametre tahminleri elde edilmiştir. Başlangıç değerleri için tanımlayıcı istatistiklerden de faydalanılabilir. İki tepeli von Mises karma dağılımının parametre tahmini için en çok olabilirlik yöntemine bağlı olarak yazılan R kodu:

```
library(maxLik)
library(circular)
theta=rad(dataset) #Dairesel veri seti theta, rad: radyan
#başlangıç değerler, p0,mu10,mu20, k10 ve k20
# von Mises olasılık yoğunluk fonksiyonu dvonmises
logLikMix <- function(param) {
p<- param[1]
if (p < 0 || p> 1)
return(NA)
mu1 <- param[2]
mu2 <- param[3]
kappa1 <- param[4]
kappa2 <- param[5]
ll <- log(p * dvonmises(theta, mu1,kappa1) + (1 - p) * dvonmises(theta, mu2,kappa2))
}
summary(m1 <- maxLik(logLikMix, start = c(rho = rho0,
mu1 = mu10, mu2 = mu20,k1=k10,k2=k20)))
```

Benzer şekilde, diğer karma ve genelleştirilmiş von Mises dağılımı için de basit modifikasyonlarla parametre tahmini için aynı kod kullanılabilir.

Karma dağılımlar için diğer parametre tahmin metotları da mevcuttur.[4,5] Bu dağılımlar için karma dağılım sayısı bilinmediği durumlarda Dirichlet süreci kullanılarak Bayesci yarı parametrik yöntemlerde önerilmiştir.[15,16]

BULGULAR

Bu bölümde, tek tepeli, iki-tepeli karma dairesel dağılımlar ve genelleştirilmiş von Mises dağılımlarını kullanarak, ilk önce DASSD[1] protein veri tabanındaki Alanin amino asitler dizisi (Ala-Ala-Ala) için dihedral açısı ψ, daha sonra ise su kaplumbağalarının kumsaldan dönüş yönleri verileri analiz edildi. Bu veri setlerinin analizinde açı birimi derece yerine radyan kullanılmıştır, yani derece birimiyle ifade edilen açı değerleri $\pi/180$ ile çarpılmıştır.

DİHEDRAL ψ AÇILARIN DAİRESEL ANALİZİ

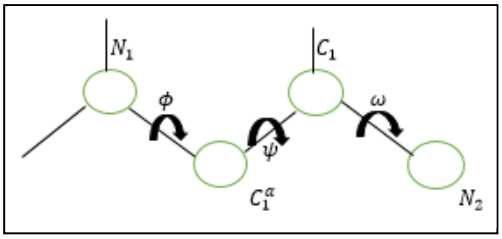
Dihedral açılar proteinlerin molekülünün geometrik yapılarından hesaplanan bağ açılarıdır.

Özetle, bir protein molekülü amino asitlerin dizisinden oluşur. Amino asitlerde karbon, hidrojen, azot gibi atomların birleşmesiyle oluşan moleküllerdir. Proteinler yapı olarak 3 kısmı ayrılır ve proteinlerin katlanmamış yapısına birincil yapı, protein zincirleri arasında iç bağlantılar oluşması sonucu oluşan sarmallara ikincil yapı, bu sarmallar arasında oluşan bağlantılar sonucu oluşan yapıya ise üçüncül yapı denir. Birincil yapının en genel tanımı ise polipeptid zincirdeki amino asit dizisi olarak tanımlanır.

Bir polipeptid zinciri ise aşağıdaki gibi atomlar dizisinden oluşur.

$$N_1 - C_1^{\alpha} - C_1 - N_2 - C_2^{\alpha} - C_2 - \ldots - N_p - C_p^{\alpha} - C_p$$

Bir proteinin yapısı ise üç dihedral açısıyla açıklanabilir. Bunlar sırasıyla ϕ, ψ ve ω' dır. ϕ açısı $-N_1 - C_1^{\alpha}$ bağı arasındaki açı, ,ψ açısı ise $-C_1^{\alpha} - C_1$ bağı arasındaki açı ve ω açısı ise $-C_1 - N_2$ bağı arasındaki açıyla Şekil 4'deki gibi gösterilir. Burada zincir üzerindeki aminoasitlerin arasındaki açılardan faydalanılarak proteinlerin ikincil yapısı veya katlanma şekilleri belirlenebilir. Genellikle ψ ve ω açı verilerinin tepe noktaları birden fazla gözlenebilir. Bu çalışmada, özel olarak DASSD[1] protein veri tabanındaki Ala-Ala-Ala protein dizisindeki ψ dihedral açısı dairesel dağılımlarla analiz edildi. Bu veri setinde



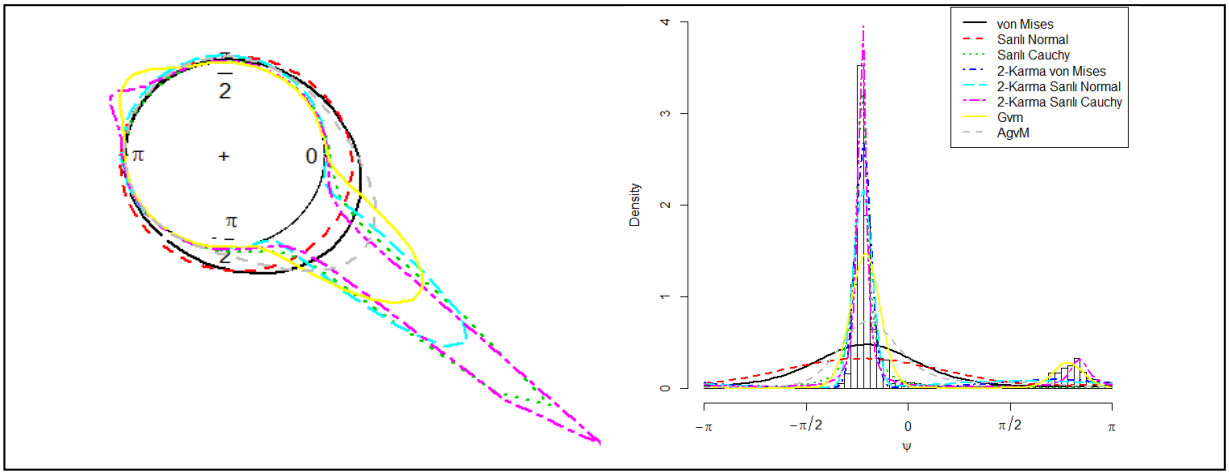
ŞEKİL 4: Bir protein molekülünün yapısı ve dihedral açıları.

875 gözlem değeri mevcuttur ve bu gözlemlerin tanım aralıkları $\psi \in [-\pi, \pi)$'dir. Bu dairesel veriyi analiz etmek için, von Mises, sarmal Cauchy, sarmal Normal, karma von Mises, karma sarmal Cauchy, sarmal Normal, genelleştirilmiş von Mises, ve üç-parametreli asimetrik von Mises, dağılımları kullanılmıştır.

Modellerin parametre tahminleri, Akaike ve Bayesci bilgi kriterleri (AIC, BIC) Tablo 1'de gösterilmiştir. Bu sonuçlara göre, 2-tepeli karma sarmal Cauchy dağılımı bu veri seti için en küçük Akaike ve Bayesci bilgi kriterlerine sahip olup, en iyi uyumu veren dağılım olarak belirlenmiştir (Tablo 1). ψ açıları için kullanılan dairesel modellerin uyumları birim daire üzerinde Şekil 5a'da gösterilmiştir. Lineer histogram üzerindeki modellerin uyumları ise Şekil 5b'de verilmiştir. İki-tepeli karma sarmal Cauchy olasılık yoğunluk fonksiyonun lineer histogram üzerinde de iyi bir uyum verdiği görülmüştür (Şekil 5b).

TABLO 1: ψ açıları için, en çok olabilirlik yönteminden elde edilen dairesel dağılımların parametre tahminleri, Akaike ve Bayesci bilgi kriterleriyle (AIC, BIC) model seçimi.

Dağılımlar	Parametre tahminleri					AIC	BIC
Tek-tepeli							
	μ	κ					
$h_{vM}(\theta_i;\mu,\kappa)$	-0.65	1.74				2371.30	2380.35
	μ	ρ					
$h_{sN}(\theta_i;\mu,\rho)$	-0.7	0.46				2710.35	2719.90
$h_{sC}(\theta_i;\mu,\rho)$	-0.71	0.91				864.38	873.93
İki tepeli							
	p	μ_1	μ_2	κ_1	κ_2		
$ph_{vM}(\theta_i;\mu_1,\kappa_1)+(1-p)h_{vM}(\theta_i;\mu_2,\kappa_2)$	0.78	-0.69	2.24	73.85	1.34	467.26	491.13
	p	μ_1	μ_2	ρ_1	ρ_2		
$ph_{sN}(\theta_i;\mu_1,\rho_1)+(1-p)h_{sN}(\theta_i;\mu_2,\rho_2)$	0.77	-0.70	1.83	0.99	0.49	475.64	499.51
$ph_{sC}(\theta_i;\mu_1,\rho_1)+(1-p)h_{sC}(\theta_i;\mu_2,\rho_2)$	0.84	-0.71	2.54	0.94	0.85	231.71	255.58
$h_{gvM}(\theta_i;\mu_1,\mu_2,\kappa_1,\kappa_2)$	μ_1	μ_2	κ_1	κ_2			
	-0.45	-0.67	0.85	4.97		738.24	757.33
$h_{agvM}(\theta_i;\mu,\kappa_1,\kappa_2)$	μ	κ_1	κ_2				
	-0.09	1.90	-1.00			1698.5	1712.82



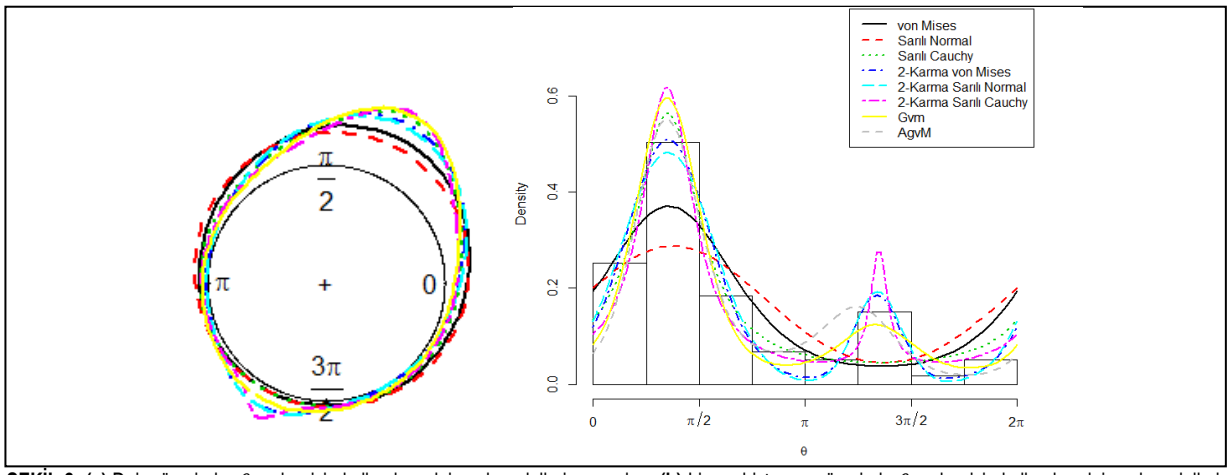
ŞEKİL 5: (a) Daire üzerinde ψ açıları için kullanılan dairesel modellerin uyumları. **(b)** Lineer histogram üzerinde ψ açıları için kullanılan dairesel modellerin uyumları.

TABLO 2: θ açıları için, en çok olabilirlik yönteminden elde edilen dairesel dağılımların parametre tahminleri ve Akaike ve Bayesci bilgi kriterleriyle (AIC, BIC) model seçimi

Dağılımlar	Parametre tahminleri					AIC	BIC
Tek-tepeli							
	μ	κ					
$h_{vM}(\theta_i;\mu,\kappa)$	1.12	1.15				243.08	247.74
	μ	ρ					
$h_{sN}(\theta_i;\mu,\rho)$	1.18	0.38				254.16	258.82
$h_{sC}(\theta_i;\mu,\rho)$	1.11	0.56				230.48	235.14
İki tepeli							
	p	μ_1	μ_2	κ_1	κ_2		
$ph_{vM}(\theta_i;\mu_1,\kappa_1)+(1-p)h_{vM}(\theta_i;\mu_2,\kappa_2)$	0.84	1.11	4.21	2.62	8.45	220.82	232.47
	p	μ_1	μ_2	ρ_1	ρ_2		
$ph_{sN}(\theta_i;\mu_1,\rho_1)+(1-p)h_{sN}(\theta_i;\mu_2,\rho_2)$	0.83	1.11	4.22	0.79	0.94	221.64	233.29
$ph_{sC}(\theta_i;\mu_1,\rho_1)+(1-p)h_{sC}(\theta_i;\mu_2,\rho_2)$	0.88	1.10	4.24	0.63	0.86	224.10	235.75
$h_{gvM}(\theta_i;\mu_1,\mu_2,\kappa_1,\kappa_2)$	μ_1	μ_2	κ_1	κ_2			
	1.18	1.08	0.79	0.96		222.5	231.82
$h_{agvM}(\theta_i;\mu,\kappa_1,\kappa_2)$	μ	κ_1	κ_2				
	1.74	0.88	-1.00			224.76	231.75

SU KAPLUMBAĞALARININ KUMSALDAN DÖNME AÇILARININ DAİRESEL ANALİZİ

Bu veri seti, Gould tarafından gözlenmiş ve 76 su kaplumbağasının kumsala yumurtalarını bıraktıktan sonra kumsaldan dönme açılarından oluşmaktadır.[2] Bu veri setindeki gözlemlerin tanım aralıkları ise $\theta\epsilon$ [0, 2π)'dir. Bu veri setinin modellenmesinde, tek tepeli, iki-tepeli karma ve genelleştirilmiş von Mises dairesel dağılımları kullanılmıştır. Tablo 2'de bu çalışmada kullanılan dairesel dağılımların parametre tahminleri ve Akaike ve Bayesci bilgi kriterleri verilmiştir. İki-tepeli karma von Mises dağılımı ve asimetrik genelleştirilmiş von Mises dağılımları en küçük Akaike ve Bayesci bilgi kriterlerine sahip olup, en iyi uyumları vermiştir (Tablo 2). Su kaplumbağalarının dönüş yönleri verisi için kullanılan dairesel modellerin uyumları, bir birim daire üzerinde, Şekil 6a'da gösterilmiştir. Lineer histogram üzerinde de iki-tepeli karma von Mises ve asimetrik genelleştirilmiş von Mises dairesel modellerin iyi bir uyum verdiği Şekil 6b'de görülmektedir.



ŞEKİL 6: (a) Daire üzerinde θ açıları için kullanılan dairesel modellerin uyumları. **(b)** Lineer histogram üzerinde θ açıları için kullanılan dairesel modellerin uyumları.

SONUÇ VE ÖNERİLER

Biyoloji, tıp, çevrebilimi gibi çeşitli alanlarda dairesel verilerin modellenmesi önemlidir. Genellikle, araştırmalarda ortaya çıkan dairesel verilerde ise, tek bir dairesel dağılımla modelleme yeterli olmamaktadır. Çoğu durumlarda, iki tepeli dairesel verilerle karşılaşılmaktadır. Bu makalede, iki tepeli karma ve genelleştirilmiş von Mises dağılımları, proteinlerdeki dihedral açıları ve su kaplumbağalarının kumsaldan dönüş yönleri verileriyle araştırılmış ve parametre tahmin edicileri elde etmek için uygulanan iteratif yöntemler, R programında yazılan kodlarla elde edilmiştir. Bununla birlikte, kullanılan tek tepeli, iki tepeli karma ve genelleştirilmiş von Mises dairesel dağılımları, birbirleriyle Akaike ve Bayesci bilgi kriterleriyle karşılaştırılmıştır. Sonuç olarak dairesel verilerde dağılımın bir veya birden fazla tepe noktasında aşırı bir yoğunlaşma olduğunda, iki-tepeli karma von Mises ve genelleştirilmiş von Mises dağılımları, modelleme olarak tercih edilmeyebilir. Örneğin, dihedral açıları analiz ederken, von Mises dağılımlarının genelleştirmeleri yerine iki-tepeli karma sarmal Cauchy dağılımı ile daha iyi bir modelleme elde edilmiştir (Tablo 1). Dairesel dağılımların tepe noktalarında aşırı bir yoğunlaşma gözlenmediğinde ise genelleştirilmiş von Mises dağılımları dairesel verileri modellemek için önerilebilir. Karma dairesel dağılımlara göre, genelleştirilmiş von Mises dağılımlarının parametre sayısının az olması tepe noktalarında aşırı yoğunlaşma olmayan dairesel veriler için model karşılaştırmalarında bir avantaj olarak değerlendirilebilir. Son olarak model seçimi için, klasik yöntemler olan Akaike ve Bayesci bilgi kriterleri yerine, dairesel verilere özgü yeni bilgi kriterleri, dairesel veriler için ilgi çekici araştırma konuları olabilir.

KAYNAKLAR

1. Dayalan S, Gooneratne ND, Bevinakoppa S, Schroder H. Dihedral angle and secondary structure database of short amino acid fragments. Bioinformation 2006;1(3): 78-80.
2. Stephens MA. Tests for von Mises distribution. Biometrika 1969;56(1):149-60.
3. Mardia KV. Statistics of directional data. J R Stat Soc Series B (Methodological) 1975;37(3):349-93.
4. Chang-Chien SJ. Hung WL, Yang MS. On mean shift based clustering for circular data. Soft Comput 2012;16(6):1043-60.
5. Hung, WL, Chang-Chien, SJ, Yang MS. Self updating clustering algorithm for estimating parameters in mixtures of von Mises distributions. Journal of Applied Statistics 2012;39(10):2259-74.
6. Cox DR. Contribution to discussion of Mardia. J R Stat Soc Series B 1975;37(3): 380-1.
7. Yfantis EA, Borgman LE. An extension of the von Mises distribution. Communications in Statistics-Theory and Methods 1982;11(15):1695-6.

8. Kim S, SenGupta A. A three parameter generalized von Mises distribution. Stat Pap 2013;54(3):685-93.
9. Gatto R, Jammalamadaka S. The generalized von Mises distribution. Stat Methodology 2007;4(3):341-53.
10. Fisher NI. Analysis of a single sample of data. Statistical Analysis of Circular Data. 1st ed. New York: Cambridge University Press; 1993. p.59-102.
11. Mardia KV, Jupp PE. Point estimation. Directional Statistics. 2nd ed. London: John Wiley & Sons Ltd; 1999. p.83-90.
12. Jammalamadaka SR, Sengupta A. Estimation of parameters. Topics in Circular Statistics. 1st ed. New York: World Scientific Press; 2001. p.85-97.
13. Abramowitz, M. and Stegun, IA. Bessel functions of integer order. Handbook of mathematical functions; with formulas, graphs and mathematical tables. New York: Dover Publications; 1965. p.358-89.
14. Henningsen A, Toomet O. maxLik: a package for maximum likelihood estimation in R. Computational Statistics 2011; 26(3):443-58.
15. Bhattacharya S, SenGupta A. Bayesian analysis of semiparametric linear-circular models. Journal of Agricultural Biological and Environmental Statistics (JABES) 2009;14(1):33-65.
16. Nuñez-Antonio G, Ausín M, Wiper MP. Bayesian nonparametric models of circular variables based on Dirichlet process mixtures of normal distributions. Journal of Agricultural, Biological and Environmental Statistics (JABES) 2015;20(1):47-64.